A COMPARATIVE EVALUATION OF THE EFFECTS OF RATER PARTICIPATION AND RATER TRAINING ON CHARACTERISTICS OF EMPLOYEE PERFORMANCE APPRAISAL RATINGS AND RELATED MEDIATING VARIABLES

A DISSERTATION

Presented to

The Faculty of the Division of Graduate Studies

By

William Irvin Sauser, Jr.

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy in the School of Psychology

Georgia Institute of Technology

December, 1978

A COMPARATIVE EVALUATION OF THE EFFECTS OF

RATER PARTICIPATION AND RATER TRAINING ON

CHARACTERISTICS OF EMPLOYEE PERFORMANCE APPRAISAL

RATINGS AND RELATED MEDIATING VARIABLES

Approved:

Edward H. Loveland, Ph.D., Chairman

Scarvia B. Anderson, Ph.D.

Sherman F. Dallas, Ph.D.

Morris Mitzner, Ph.D.

Charles V. Riche, Ph.D.

Date approved by Chairman

ACKNOWLEDGMENTS

The design and enactment of a major research project and the writing of a doctoral dissertation describing that project are tasks which cannot be accomplished without advice and support from mentors, colleagues, and friends. I thank the members of my advisory committee--Doctors Scarvia Anderson, Sherman Dallas, Morris Mitzner, and Charles Riche--not only for providing helpful guidance but also for sustaining me through the completion of my graduate studies. A special expression of gratitude is reserved for Dr. Edward Loveland, my academic and personal advisor and dissertation direction, for contributing in so many ways to this project and to my professional education.

A vote of appreciation is due the faculty and graduate students of the School of Psychology at the Georgia Institute of Technology for their encouragement and assistance throughout my graduate training. I also thank my colleagues in the Department of Psychology at Auburn University for their patience, advice, and support during the completion of this dissertation.

I am heavily indebted to my friends who gave freely of their time to assist me in coding and tabulating data, punching and sorting cards, and assembling and evaluating questionnaire items: Jay Anderson, Phil and Jeanne Lemkau, Bob and Susan Pond, Tom and Trudy Stutzman, and "C.J." Wilson. I also thank Pat Watson for her excellent work in typing the final manuscript.

This dissertation is dedicated to three persons who have meant so much to me--my splendid wife, Lane, my stalwart father, William, and the memory of my beloved mother, Kathryn.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

LIST OF TABLES vi

LIST OF ILLUSTRATIONS. viii

SUMMARY. ix

Chapter

I. LITERATURE REVIEW 1

The Concept of the Criterion
"Hard" Criteria
"Soft" Criteria
Rating "Errors"
Rating Theory
A Taxonomy of Sources of Variance and Error in Ratings
Approaches to Reducing Error in Ratings

II. STATEMENT OF THE PROBLEM. 52

A Hypothetical Model of the Effects of Training and Participation
Variables Under Consideration
Plan of Analysis and Hypotheses

III. METHOD OF INVESTIGATION 65

Subjects
Instruments and Materials
Treatment Conditions and Procedure

IV. RESULTS 93

Study One
Study Two
Study Three
Study Four
Study Five

TABLE OF CONTENTS (Cont'd)

Page

Study Six
Study Seven
Study Eight
Study Nine
Study Ten

V. DISCUSSION AND IMPLICATIONS 148

Implications for the Hypothetical Model
Limitations of the Investigation
Suggestions for Future Research

APPENDICES 167

A. Informed Consent Forms
B. Attitude Scales
C. Knowledge Scale
D. Criterion Scale
E. Critical Incident Reporting Forms
F. Behaviorally Anchored Rating Scales
G. Simulated Professors
H. Rater Training Program

BIBLIOGRAPHY 219

Reference Notes
References

VITA 249

LIST OF TABLES

Table		Page
1.	A Taxonomy of Sources of Variance and Error in Ratings	28
2.	Subject Participation in the Experiment Proper	67
3.	Subject Participation in the Questionnaire Development Studies.	68
4.	Means, Standard Deviations, and Reliability Estimates for the Knowledge Scale.	76
5.	Intercorrelations and Reliability Coefficients of the Attitude, Knowledge, and Criterion Scales	78
6.	Definitions of the Five Categories of College Classroom Teaching Behavior.	80
7.	Scale Values in the Simulated Professor x Category Matrix.	86
8.	Study One ANOVA Table--All Subjects.	95
9.	Study One ANOVA Table--Participant Subjects Only	96
10.	Study One ANOVA Table--Non-Participant Subjects Only	97
11.	Study One ANOVA Table--Trained Subjects Only	98
12.	Study One ANOVA Table--Untrained Subjects Only	99
13.	Study Two MANOVA and ANOVA Tables.	104
14.	Cell Means of Study Two Elevation Scores	106
15.	Study Three MANOVA and ANOVA Tables.	110
16.	Cell Means of Study Three Variance Scores.	112
17.	Mean Category Intercorrelations for Each Simulated Professor Within Each Cell of the Design	115

LIST OF TABLES (Cont'd)

Table		Page
18.	Summary of Study Four χ^2 Tests	116
19.	Study Five Reliability and Validity Scores	118
20.	Study Five ANOVA Tables.	120
21.	Study Five Participation x Category Cell Means for Intra-Class and One-Rater Validity Scores	124
22.	Study Six ANOCOV Table	127
23.	Study Seven ANOCOV Tables.	132
24.	Study Seven Adjusted Cell Means.	133
25.	Study Eight ANOVA and ANOCOV Tables.	136
26.	Study Ten MANOCOV and ANOCOV Tables.	140

LIST OF ILLUSTRATIONS

Figure — Page

1. A Hypothetical Model of the Effects of Training and Participation 53

2. Procedural Plan of the Experiment 89

3. Distribution of Variance Scores Analyzed in Study Three 109

4. A Revised Hypothetical Model of the Effects of Training and Participation. 157

SUMMARY

A review of literature dealing with the use of performance appraisal ratings as employee evaluation criteria led to the formulation of a model of the rating process. This model views the obtained rating as determined by aspects of (a) rater, (b) rating instrument, (c) rating context, (d) temporal situation, (e) ratee, (f) ratee's performance, (g) behavioral characteristic being evaluated, (h) behavioral context, and (i) interactions among these factors. All factors save (f) were classified as sources of psychometric error in ratings. Several treatments intended to reduce the influence of such psychometric error were identified; two of these--rater training and rater participation in scale construction--were selected for further investigation. A hypothetical model of the effects of Participation and Training on psychometric characteristics of ratings was formulated. This model suggests that the effects of the two treatments are mediated by changes in attitude toward and knowledge about the performance rating process. Sixteen families of hypotheses generated from this model were chosen for empirical investigation.

Ninety-six undergraduate students taking courses in psychology at Auburn University were randomly assigned to four cells of the experimental design: (a) Both Participation and Training, (b) Participation Only, (c) Training Only, and (d) Neither Participation nor Training. All subjects were pre-tested on Attitude and Knowledge.

Subjects in cells (a) and (b) participated in the construction of a set of behaviorally anchored rating scales (BARS) for measuring five aspects of college classroom teaching performance, while subjects in cells (c) and (d) performed a control task. Later, subjects in cells (a) and (c) were exposed to a rater training program, while subjects in cells (b) and (d) performed a control task. All subjects then evaluated five standardized simulated professors using the BARS. These "simulated professors" consisted of short biographical descriptions followed by behavioral diaries containing scaled incidents obtained during the BARS construction process. The diaries were prepared such that "true scores" of the simulated professors were known. After all ratings were completed, subjects were administered post-tests of Attitude and Knowledge. The entire experiment was conducted within a five-week period.

Results of ten analytic studies provided limited support for the hypothetical model. Training was found to significantly reduce overall elevation (leniency error) and unwanted variance attributable to Professors (consensual halo error). Participation was found to significantly reduce consensual halo error, to significantly increase the proportion of variance in ratings attributable to the Professor x Category interaction (discriminant validity)., and to significantly increase estimates of intraclass and one-rater reliability and validity in the sets of ratings. Participation significantly affected Attitude, although in the direction opposite from that predicted, while Training significantly increased Knowledge.

Little evidence for a Participation x Training interaction was produced; yet this effect did significantly influence posttest Knowledge scores and the dispersion of ratings per category. The significant effects of Participation and Training on psychometric characteristics of the ratings persisted when the effects of the treatments on changes in Attitude and Knowledge were held constant through covariance analysis, suggesting that additional variables mediate the treatment-rating characteristic relationship.

Implications of these findings for subjective measurement of individual differences were considered, and a revised hypothetical model was presented. Specific suggestions for a systematic investigation of the rating process were offered.

CHAPTER I

LITERATURE REVIEW

The Concept of the Criterion

The Need for Criteria of Performance

Important decisions regarding such matters as the effectiveness of training, counseling, and educational programs in upgrading employee and organizational unit effectiveness; the utility of tests, interviews, and other devices for selecting and classifying personnel; the influence of environmental/situational conditions and organizational change on employee and unit functioning; the relative levels of performance of employees being considered for promotion, demotion, transfer, or dismissal; equitable compensation for job performance; and the relative effectiveness of various work methods, schedules, and conditions are among those being dealt with daily by decision-makers in today's modern organizations, whether they be industrial, commercial, military, governmental, political, educational, religious, recreational, or service-oriented. In order to make intelligent decisions with regard to these issues, it is essential for the decision-makers to have accurate criteria of personnel performance (Baylie, Kujawski, & Young, 1974, pp. 162-163; Ghiselli & Brown, 1955, p. 60).

In addition to their obvious pragmatic importance to organizational decision-makers, criteria have legal implications as well: Employee performance criteria, and the decisions based on them, must

meet rigorous government standards (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978) intended to alleviate discrimination against minority groups and women. There are professional/ethical standards for criteria as well (American Psychological Association, 1974; Division of Industrial-Organizational Psychology, 1975).

Accurate criteria are of the utmost concern to industrial-organizational psychologists (Blum & Naylor, 1968, p. 174; Campbell, Dunnette, Lawler, & Weick, 1970, p. 101). Blum and Naylor (1968) recognize the overriding importance of criteria for the science of industrial-organizational psychology:

> The criterion is basic to all measurement in industrial psychology. To overstate its importance would be literally impossible. Without adequate criteria, industrial psychology is ineffective and ceases to be a science. In other words, the magnitude of the contribution of industrial psychology is completely determined by the adequacy of the criterion measures evolved. (p. 174)

Of course, the importance of the criterion for psychology is not limited to its industrial-organizational application, as Weitz (1961, 1964) has so elegantly demonstrated in the domain of experimental psychology, and Krasner (1971), Mischel (1968), and Wiggins (1973) have emphasized for personality and clinical psychology.

The "Criterion Problem"

Naturally, such an important concept as the criterion could be expected to give rise to a considerable body of discussion and debate. This expectation has certainly been fulfilled; some of the more important theoretical treatments of the "criterion problem" include those of Astin (1964), Bass (1962), Bechtoldt (1947), Bellows (1941), Brogden

and Taylor (1950b), Dunnette (1963), Fiske (1951), Ghiselli (1956), Ghiselli and Haire (1960), Guion (1961), James (1973), Jenkins (1946), Nagle (1953), Prien (1966), Ronan and Prien (1966), Schmidt and Kaplan (1971), Smith (1976), Toops (1944), Van Dusen (1947), Wallace (1965), Weitz (1961), and Wherry (1957).

While there may be disagreement over the optimal levels of dimensionality, complexity, dynamism, objectivity, immediacy, or ultimacy of criteria (Smith, 1976), theorists do agree that measures of criteria must possess two essential characteristics to be useful: They must be valid and they must be reliable (Ghiselli & Brown, 1955, p. 60). Other "necessary and/or desirable" characteristics of criterion measures have been listed; these include realism, representativeness, relationships with other criterion measures, acceptability to the job analyst, acceptability to management, consistency from one situation to another, predictability (Bellows, 1961, pp. 361-364); inexpensiveness, understandability, measurability, relevance, freedom from contamination and bias, discrimination (Blum & Naylor, 1968, p. 182); and practicality (Smith, 1976, p. 746). Jenkins (1946) and Wherry (1957) warn psychologists not to be unduly concerned with some of these "desirable" characteristics of criterion measures, for adequacy and validity are far more important characteristics than are expediency or convenience.

Three Criterion Constructs and Their Measurement

The term "validity" is often reserved to describe predictors or independent variables. Criteria of performance are typically dependent variables or measures which are being predicted. The term relevance

is used to refer to the "validity" of a performance criterion measure. Blum and Naylor (1968), drawing from the work of Brogden and Taylor (1950b) and Nagle (1953), define relevance and related concepts as follows:

> One way of viewing criteria is in terms of actual and ultimate measures of success. By definition, whatever measure of success one actually uses is the actual criterion. The ultimate criterion, on the other hand, is a theoretical and ideal criterion that usually exists only in the psychologist's mind. It is the "true" criterion of success, while our actual criterion is the measure we have been forced to adopt simply because we can do no better. . . .
>
> The three major constructs of a criterion are deficiency, relevance, and contamination. . . .Criterion deficiency is the degree to which our criterion is lacking certain variance necessary to the ultimate criterion. . . .Criterion relevance is the degree to which the actual criterion overlaps or corresponds to the true criterion. . . .Criterion contamination is that variance in the actual criterion which is unrelated to the ultimate criterion. (pp. 176-177)

Blum and Naylor (1968, p. 177) further classify contamination as either systematic (bias) or random (error).

Defining criterion relevance in terms of the overlap of the actual criterion with a hypothetical, unmeasurable ultimate criterion points up a major difficulty in evaluating criteria: "There is no way to validate the criterion objectively since there is no objective, external basis for use as a standard with which it may be compared" (Bellows, 1941, pp. 506-507). Logical evaluation and judgment regarding relevance is often cited as the only method whereby the adequacy of a criterion measure can be evaluated (Astin, 1964; Brogden & Taylor, 1950b; Guion, 1961). This logical evaluation and judgment process, commonly referred to as content validation (American Psychological Association,

1974; Cronbach, 1970, pp. 145-148; Division of Industrial-Organizational Psychology, 1975), is exemplified in the criterion development process outlined by Nagle (1953, p. 285).

In addition to the content validation process, the construct validation approach (Cronbach, 1970, pp. 142-144) has been recommended for evaluating the relevance of criterion measures by some theorists (Guion, 1961; James, 1973). Specific construct validation methods recommended for criterion evaluation include the multitrait-multimethod matrix (Campbell & Fiske, 1959; Dickinson & Tice, 1973; Kavanagh, MacKinney, & Wolins, 1971), the multitrait-multirater approach (Lawler, 1967; Zedeck & Baker, 1972), factor analysis (Ewart, Seashore, & Tiffin, 1941; Grant, 1955; Ronan, 1963a, 1963b), and multimethod factor analysis (Jackson, 1969). While these methods may certainly be useful for examining and evaluating criterion measures, each of them contains as an integral component comparison of relationships among a set of criterion measures, each of which may or may not be relevant. Therefore, unless each of the measures in this set has been examined for contamination and deficiency through the content validation approach, the application of these sophisticated construct validation techniques to them may be improper.

Smith (1976) identifies and describes two types of criteria, "The 'hard' criteria obtained from organizational records such as absences and turnover, and the 'soft' criteria obtained from ratings. The first maintains the appearance of objectivity; the second is frankly judgmental" (p. 753). Each of these two types of criteria is briefly reviewed below.

"Hard" Criteria

Examples of "Hard" Criteria

Examples of "hard" criteria include tardiness (Mueser, 1953), absences (Huse & Taylor, 1962; Kerr, Koppelmeir, & Sullivan, 1951; Metzner & Mann, 1953), accidents (Daniels & Edgerton, 1954; Whitlock, Clouse, & Spencer, 1963), turnover (reviewed comprehensively by Schuh, 1967), sales (Rush, 1953; Taylor, Schneider, & Symons, 1953; Weitz & Nuckols, 1953), length and shape of learning curves (Lefkowitz, 1970; Taylor & Smith, 1956), job level and promotions (Bentz, 1968; Henry, 1948), salary (Bingham & Davis, 1924; Gifford, 1928; Hulin, 1962), and production (Guion, 1965, pp. 93-94; Smith & Gold, 1956).

The Problems of Contamination and Deficiency

While the "so-called objective" (Smith, 1976, p. 753) criteria listed above may appear face-valid as valuable measures of employee performance, they are often heavily contaminated by factors partially or completely beyond the employees' control, such as working conditions and units of production (Bellows, 1941; Brogden & Taylor, 1950b; Ronan & Prien, 1966; Toops, 1944; Ronan, Note 1). Most of the criterion measures listed above are contaminated by uncontrollable situational factors. Such contaminants have been identified for absence rate (Behrend, 1953), turnover (Behrend, 1953; Stark, 1959; Tiffin & Phelan, 1953), accidents (Smith, 1976, p. 754) and accident rates (Ghiselli & Brown, 1955, p. 344), job level, promotions, and salary (Hulin, 1962; Smith, 1976, pp. 756-757), sales criteria (Dorcus, 1940), quantity of output measures (Bellows, 1941, p. 503; Toops, 1944), and measures of

quality of performance (Flanagan, 1948, pp. 126-128). Ronan (Note 1) cautions that insidious factors such as family and ethnic customs (Viteles, 1932, pp. 213-214), informal work group rate restrictions (Lupton, 1963; Viteles, 1932, pp. 560-565, 1953, pp. 45-61; Whyte, 1955), and the complex interplay of attitudes, perceptions, and opinions (Patchen, 1970; Viteles, 1936) can also seriously affect these "so-called objective" measures of personnel performance.

As Campbell et al. (1970) point out, many objective criteria are not only contaminated but also "seriously deficient in that they include only a few rather than all or many of the behavioral elements making up a job" (p. 107). Using such criteria could place undue emphasis on some aspects of performance, at the expense of other, perhaps equally important aspects. The fascinating experiences reported by Blau (1963) show what can happen when deficient criteria are employed.

Attempts to Avoid Contamination and Deficiency

Recognizing the importance of the problems of contamination and deficiency, a number of investigators are attempting to develop "hard" criteria which are representative of the complete job and are not so easily influenced by situational and other biasing factors. Examples of such attempts are the "psychometric approach to job performance," spearheaded in the literature by Ronan and his colleagues (Atlanta Regional Commission, 1974; Ronan, Anderson, & Talbert, 1976; Talbert, Carroll, & Ronan, 1976); the simulation technique (Besnard & Briggs, 1967; Viteles, 1945); and the situational tests (Flanagan, Fiske, Bass, Carter, Kelly, & Weislogel, 1954) incorporated into many assessment

centers (Bray, 1964; Bray & Campbell, 1968; Bray & Grant, 1966; Byham, 1970; Byham & Thornton, 1970; Finkle, 1976), including the "in-basket test" (Frederiksen, Saunders, & Wand, 1957; Lopez, 1966) and the "leaderless group discussion" (Bass, 1954). However, many of the techniques included in assessment centers typically rely heavily on subjective judgment on the part of the assessors.

Other techniques which combine judgment, observation, and record-keeping include Brogden and Taylor's (1950a) "dollar criterion," Flanagan's (1954) "critical incident technique" and its derivative reporting forms and checklists (Campion, 1972; Flanagan, 1949; Flanagan & Burns, 1955; Rambo, 1958; Ronan, 1972), and various anecdotal report forms and files (Guion, 1965, pp. 468-469; Smith, 1976, p. 752). Smith (1976) classifies these latter approaches as measures of behaviors as opposed to results, and implies that they are subject to bias through both situational and judgmental factors. Furthermore, she indicates that all of the "hard" criteria are subject to some degree of judgmental bias: "The so-called hard criteria all involve some subjective components. Human judgment enters into every criterion from productivity to salary increases" (p. 757).

"Soft" Criteria

Judgmental Methods

In many cases it is impossible, or prohibitively expensive, to identify and/or construct "bias free" objective criteria (Bellows, 1961, p. 351). As an alternative to this approach to criterion measurement, researchers and practitioners have typically turned toward the development

of subjective evaluation devices modeled after several of the psychophysical scaling methods (Guilford, 1954; Jones, 1974; Torgerson, 1958). Gulliksen (1958) indicates that psychophysics, as developed by Fechner (1860), originally included only the measurement of sensory attributes and the quantification of perception, in order to correlate the psychological scales with physical measurements of the stimuli. However, according to Gulliksen, Thurstone's (1927) monumental work "pointed out that many of these 'psychophysical' scaling methods could be used for accurate measurement of psychological attributes of stimuli which had no relevant measurable physical correlate" (p. v). One such stimulus, apparently, is employee performance.

The ranking method and the method of paired comparisons (among employees) have been used in the employee performance appraisal context (Blum & Naylor, 1968, pp. 206-209; Ghiselli & Brown, 1955, pp. 96-103; Lawshe, Kephart, & McCormick, 1949). Other innovative applications of scaling techniques to performance appraisal include the use of scaled check-list items (Ferguson, 1947; Knauft, 1948; Uhrbrock, 1950, 1961), forced-choice scales (Berkshire & Highland, 1953; Kay, 1959; Obradovic, 1970; Richardson, 1949; Sisson, 1948; Taylor, Schneider, & Clay, 1954; Taylor et al., 1953; Travers, 1951; Zavala, 1965), "man-to-man" scales (Guilford, 1954, pp. 269-270; Ross, 1966), and Hartshorne and May's (1929) "portrait matching" and "guess who" techniques (see Guilford, 1954, pp. 270, 272).

Ratings: The Method of Single Stimuli

The most popular by far of the psychophysical scaling methods to be adapted to the performance appraisal process is the "method of

single stimuli" (Guilford, 1954, p. 145; Torgerson, 1958, p. 67; Volkmann, 1932; Wever & Zener, 1928), commonly referred to as the rating method.

Guilford (1936, pp. 264-265) has traced the early history of the rating method. While Galton (1883) seems to be the first researcher to employ a rating scale in a psychological problem (evaluating the vividness of images), Titchener (1909) reports that such scales had been used as early as 1805 by the British Navy to describe wind strength, and that a Washington, D.C. scientist, J. W. Osborne, developed a scale for rating temperature in 1876. Early applications of the rating method to psychological problems include evaluative judgments of the affective value of colors (Major, 1895), the humor of comics (Martin, 1905), the agreeableness of odors (Keith, 1906), and the persuasiveness of advertisements (Hollingworth, 1911). Guilford credits Pearson's (1907) seven-point scale for estimating intelligence as "the first attempt to secure ratings of human abilities. . ., an application which received its greatest impetus during the World War when psychologists were called upon to devise methods of rating the efficiency of officers" (p. 265). The literature concerning the application of the rating method to a wide variety of psychological problems has burgeoned since that time. The use of rating scales as personnel evaluation devices also has a long history. Benjamin (1952) comments: "As far back as 1916 the Lord and Taylor department store had an appraisal system comparable to many in use today" (p. 289). Indeed, Mahler (1947, p. v) claims that a personnel rating form having much in common with some forms employed today was in use over 190 years ago.

Torgerson's (1958, p. 67) excellent description of the method of single stimuli as it is commonly used in the psychophysics laboratory is also appropriate for the variety of performance rating scales commonly used in industry (Bass & Barrett, 1972, p. 219; Dunnette, 1966, p. 93; Guilford, 1954, pp. 265-266; McCormick & Tiffin, 1974, p. 196). The typical industrial-organizational application of the method differs little from Torgerson's description, except that raters are evaluating qualities of employees rather than of physical stimuli and that repeated judgments are often the exception rather than the rule.

Popularity of the Rating Method

McCormick and Tiffin's (1974) statement, "Rating scales are the most widely used type of performance evaluation system" (p. 195), is strongly supported by the literature. Surveys indicate that rating scales are popular as criteria for personnel action decisions (Bass & Barrett, 1972, p. 210; Bellows, 1961, p. 376; Spriegel & Mumma, 1961) as well as for research purposes (Blum & Naylor, 1968, pp. 197-198; Guion, 1965, p. 96).

Criticism of Ratings as Performance Criteria

Despite their popularity, merit ratings have come under severe attack due to their questionable reliability and validity (Kipnis, 1960; Ronan & Schwartz, 1974; Taft, 1955; Toops, 1944). Impressive evidence can be marshalled against ratings as they are commonly obtained and used as measures of performance in industry. Interrater agreement, especially across hierarchical levels, is typically moderate to low (Besco & Lawshe, 1959; Charest, Cowart, & Goodman, 1963; Kavanagh

et al., 1971; Kirchner, 1966; Ronan & Latham, 1974; Schneider & Bartlett, 1970; Tucker, Cline, & Schmitt, 1967). "In general, over all multirater studies," state Ronan and Schwartz (1974), "it is usual to find interrater correlations on the order 0.60, but. . .correlations are generally much lower when raters are making assessments independently of each other" (p. 72). Ronan (Note 2) estimates the median interrater reliabiltiy coefficient in this latter case to be "approximately 0.25. . . [representing] an infinitesimal portion of the performance variance" (p. 7).

Based on results from some of the construct validation techniques mentioned earlier, several researchers have questioned the validity of ratings as performance criteria. Studies in which ratings are correlated with "hard" criteria typically uncover low relationships between the two types of criteria (Barrett, 1966, pp. 68-72; Gaylord, Russell, Johnson, & Severin, 1951; Kirchner, 1960; Morrison, Owens, Glennon, & Albright, 1962; Trawick & Munger, 1962; Tucker et al., 1967; Zedeck & Baker, 1972). Severin (1952), in a review of some 150 studies reporting correlations among various measures of job performance, including ratings and production records, found a median correlation of .28. Seashore, Indik, and Georgopoulos (1960) found considerable variation in interrelationships among criterion measures across a set of similar organizations. Factor analytic studies have also called the construct validity of ratings into question, since ratings generally load on factors orthogonal to "objective" measures of performance (McClelland & Rhodes, 1969; Ronan, 1963a; Rush, 1953; Turner, 1960; Taylor, Smith, Ghiselin, & Ellison, 1961).

These lines of evidence have led some writers to call for the outright rejection of ratings from consideration as performance criteria. For example, Ronan and Schwartz (1974) conclude their review as follows:

> This review has presented evidence that strongly implies that ratings of human performance fail to be replicable over independent groups and fail to be congruent with objective performance indicators. Surely, human performance should not be judged by ratings or any other tool which has properties that are not completely known or understood. (p. 79)

Arguments in Partial Defense of the Rating Method

The arguments presented in the preceding paragraphs appear to completely condemn the rating method as a performance appraisal alternative. However, several counter-arguments can be offered to partially refute some of the indicting evidence. For example, consider Bass and Barrett's (1972) rationale for the use of the correlative construct validation approach:

> Because we can seldom directly appraise an ultimate criterion, or the long-term true contribution to the organization, we do the next best thing. We select a variety of immediate measures, among them merit ratings, which we assume all relate imperfectly to the true contribution. Then, the greater the correlations among these diverse measures, the more confident we are that we are appraising (with some degree of error) the true contribution by the various measures. (pp. 213-214)

By implication, low correlations between ratings and objective measures, as documented in the preceding section, should reduce confidence in the use of ratings as criteria. Bass and Barrett's statement appears well founded; however it is commonly agreed that job performance is multidimensional (Blum & Naylor, 1968, pp. 187-189; Dunnette, 1963; Ghiselli, 1956; Guion, 1961, 1965, pp. 114-115; Roach & Wherry, 1970;

Ronan & Prien, 1966; Smith, 1976; Stogdill, Shartle, Wherry, & Jaynes, 1955; Taylor, Brice, Richards, & Jacobsen, 1964, 1965), and it is not reasonable to expect measures of such dimensions as absenteeism and tenure to correlate with rated performance on other (orthogonal) dimensions. Low intercorrelations among criteria measuring separate dimensions of performance would be a healthy sign of discriminant validity, rather than a signal to abandon the criteria. (It may be reasonable, however, to expect a high correlation, indicating convergent validity, between two criteria which purportedly measure the same dimension, such as patent disclosures and rated creativity.)

A second argument in favor of ratings is that many "objective" criteria are known to be heavily contaminated (Bellows, 1941; Brogden & Taylor, 1950b; Ronan & Prien, 1966; Toops, 1944) and may not be suitable as measures with which to evaluate ratings of performance. It would not be reasonable to expect judgmental measures of performance, which may take situational factors into account, to correlate highly with measures which are contaminated by factors completely beyond the control of the person being evaluated.

The third counter-argument relates to the problem of low intercorrelations among ratings from different sources, such as peers, subordinates, superiors, and self. Individuals within the same job context have been found to disagree in their perceptions of what specific behaviors (Borman, 1974; Crawford & Bradshaw, 1968; Ronan & Latham, 1974; Schneier & Beatty, 1978; Wiley, 1964; Zedeck, Imparato, Krausz, & Oleno, 1974; Tauscher, Note 3), characteristics and traits (Gruenfeld

& Weissenberg, 1974; Maslow & Zimmerman, 1956; Parker, Taylor, Barrett, & Martens, 1959; Prien, 1962; Stander, 1965), and organizational variables (Friedlander, 1966) influence job success. Tauscher (Note 3) has stated: "The evidence, then strongly suggests there are basic differences in the manner in which individuals or groups approach and evaluate the same job performance" (p. 10). One interpretation of Tauscher's remarks might be: Since raters cannot agree on what is important, ratings are therefore invalid. However, there is another, perhaps more subtle, interpretation: It may be that all, or most, of the viewpoints taken by the different groups of raters are relevant--that is, superiors, for example, might indeed view the job differently from incumbents, but both may be viewing separate, but equally relevant, dimensions of performance (Borman, 1974; Buckner, 1959; Campbell et al., 1970, pp. 114-116). Campbell et al. express this view as follows:

> Disagreement between different observers should not necessarily be viewed as a mark of unreliability (as tradition has so often dictated), but should instead be viewed as a possibly valid indication that differing aspects of [an incumbent's] behavior are being accurately perceived and reported. (p. 115)

Finally, as Bellows (1961, p. 351) has noted, there are some important aspects of job performance which simply cannot be assessed with objective measures. One must resort to carefully constructed subjective evaluation instruments if these aspects are to be measured at all.

While these four counter-arguments may somewhat weaken the intensity of the attack on ratings as a performance appraisal method, even the most optimistic supporters of the rating method admit that ratings, as they are typically obtained and used in industry, are laden

with serious problems (Bellows, 1961, p. 396; Campbell et al., 1970, p. 111; Rowland, 1970, p. 211; Rundquist & Bittner, 1950; Spicer, 1951). However, a number of authors, including Bass and Barrett (1972, p. 212) and Guilford (1954, pp. 297-298), hold out hope that these problems can be corrected through research and diligent application, and argue against "throwing out the baby with the bathwater" (Baylie et al., 1974, p. 186). In support of this viewpoint, there is evidence suggesting that rating devices, when carefully constructed and applied, can yield usable, relevant, acceptable measures of performance (Barrett, Taylor, Parker, & Martens, 1958; Bittner & Rundquist, 1950; Borman, 1978; Cascio & Valenzi, 1978; Ferguson, 1949b; Fogli, Hulin, & Blood, 1971; Hoyle & Arvey, 1972; Kane & Lawler, 1978; Latham & Wexley, 1977; Richardson & Kuder, 1933; Rundquist & Bittner, 1948; Smith & Kendall, 1963; Taylor, Barrett, Parker, & Martens, 1948; Taylor & Manson, 1951).

Rating "Errors"

Guilford (1954) states that "the use of ratings rests on the assumption that the human observer is a good instrument of quantitative observation, that he is capable of some degree of precision and some degree of objectivity" (p. 278). He suggests that whenever psychologists are forced to place confidence in quantitative human judgments, "we must be ever alert to the weaknesses involved and to the many sources of personal biases in those judgments" (p. 278). These personal biases are commonly referred to as "rating errors." Some of the more prevalent of these errors are discussed below.

Leniency

The error of leniency has been described simply as follows: "Ratings tend to be bunched toward the favorable end of the rating scales. The average person is rated as above average, making for a displacement of the mean, and skewness" (Smith, 1976, p. 757). Barrett (1966, pp. 23-25) and Smith (1976, p. 757) present several possible reasons for this error, including preselection of ratees, reflection of the rater's own competence, human kindness, reluctance to criticize or confront employees, and scale ambiguity. An error opposite in effect to that of leniency, severity (a low mean rating and positive skewness), can also occur (Guilford, 1954, p. 278; Smith, 1976, p. 757). Glickman (1955) and Bass (1956) have demonstrated that leniency can have detrimental effects on employee motivation and personnel administration, making it a very important error to avoid. Presumably severity can have the same effects and may be just as dangerous.

Central Tendency

This error is represented by a deviation from the expected roughly normal curve--"ratings tend to pile up in the middle of the distribution" (Smith, 1976, p. 757). Hesitation on the part of raters to give extreme judgments is the reason typically given for this error (Guilford, 1954, p. 278). While the classical definition of central tendency refers to a tight distribution around the midpoint of the scale, central tendency may be combined with the errors of severity or leniency to give a leptokurtic and skewed distribution (Smith, 1976, p. 757). Central tendency is commonly found in ratings of people or products, yet the opposite

distributional effect, bimodality (piling up items at the ends of the scales, avoiding the middle), may occur when items describing people are being rated (Cliff, 1959; Rotter & Tinkleman, 1970). Harari and Zedeck (1973), Landy and Guion (1970), and Zedeck et al. (1974) found bimodality to be a problem in the development of behaviorally anchored rating scales of job performance. Their subjects tended to rate most behavioral incidents as representative of extremely favorable or unfavorable performance, and avoided assigning neutral values to behavioral descriptors.

Halo

This pervasive error, first mentioned by Wells (1907), was given its name by Thorndike (1920), who noted that "ratings [are] apparently affected by a marked tendency to think of a person in general as rather good or rather inferior and to color the judgments of the qualities by this general feeling" (p. 25). The result of this "coloring" is that the "rating on one characteristic spills over to affect ratings on other characteristics, resulting in high intercorrelations among ratings for supposedly different characteristics or behaviors" (Smith, 1976, p. 757), thus forcing "the rating of any trait in the direction of the general impression of the individuals rated and to that extent [making] the ratings of some traits less valid" (Guilford, 1954, p. 279). Smith (1976) notes, "Halo can be either favorable or unfavorable--it merely represents the failure of the rater to differentiate" (p. 757). Guilford (1954, p. 279) believes that every judge falls victim to the halo effect, and there is indeed evidence in the literature (Borman, 1975; Brown, 1968;

Guilford, Christensen, Taaffe, & Wilson, 1962; Kornhauser, 1927; Turner, 1960; Vielhaber & Gottheil, 1965) that this error is common in ratings of performance, especially when the traits being rated are ambiguous (Symonds, 1925). While halo is typically viewed as a problem to be avoided if possible, Bingham (1939) argues that in addition to the possibility of very real relationships among traits, there is an additional correlation due to "valid" halo, "a halo which cannot and should not be eliminated because it is inherent in the nature of personality, in the perceptive process and in the very act of judgment" (p. 222).

Logical Error

Newcomb (1931) identified the logical error, an error the effect of which is not unlike that of the halo error: "Judges are likely to give similar ratings for traits that seem logically related in the minds of the raters" (Guilford, 1954, p. 279). The logical error, or implicit personality theory (Cronbach, 1955; Gage & Cronbach, 1955), has been studied by attributionists and other personality theorists (Jones, Kanouse, Kelley, Nisbett, Valins, & Weiner, 1971; Kelley, 1973; Passini & Norman, 1966, 1969; Shweder, 1975), and evidence such as that presented by Koltuv (1962) and Mulaik (1964) has led Mischel (1971) to conclude that "the factors identified by trait ratings may reflect the social stereotypes and concepts of the judges rather than the trait organization of the rated persons" (p. 141). Smith (1976) argues that the logical error and the implicit trait relationship assumptions behind it may be necessary, since "without such a set of assumptions, hardly any rating would be possible; the problem is to eliminate false generalizations or at least to systematize the assumptions held" (p. 758).

Contrast Error

Murray (1938) identified the contrast error, "a tendency for a rater to rate others in the opposite direction from himself in a trait" (Guilford, 1954, pp. 279-280). Guilford (1954, p. 280) explains the contrast error in terms of the psychoanalytical phenomena of reaction formation and projection, yet more recent research (Berkowitz, 1960; Hakel, Ohnesorge, & Dunnette, 1970; Holmes & Berkowitz, 1961) has revealed that in addition to the rater's self-concept, other stimuli, such as evaluations of other ratees, can also serve as frames of reference for contrast and comparison. Such contrast effects have been found in a variety of interpersonal evaluation situations, including assessment decisions (Rose, 1967), employment interviewing (Hakel et al., 1970; Wexley, Yukl, Kovacks, & Sanders, 1972), and social perception situations (Holmes & Berkowitz, 1961). Wexley et al. found that contrast effects were especially strong when the ratee is of "intermediate suitability" (p. 47).

Proximity Error

Stockford and Bissell (1949) discovered this error which, like the logical and contrast errors, "injects undue covariances among rated trait variables. The reason for this source of spurious correlation is the nearness in space or in time for the rating of two traits" (Guilford, 1954, p. 280). In other words, adjacent traits may be more highly correlated than more remotely separated traits on the rating scale.

Rating Theory

Guilford's Model of Ratings

Guilford (1954, pp. 280-281) presents a model equation of a rating which defines the rating of person I in trait J by rater K as a linear combination of terms representing: (a) the true value of person I in trait J, (b) rater K's leniency error, (c) rater K's halo error in connection with person I, (d) rater K's rater-trait interaction error, and (e) residual error made by rater K in rating person I. In defining his terms, Guilford classifies the contrast error as an example of (d), while the error of central tendency and the logical and proximity errors form part of (e). Guilford's model serves as the basis for a very popular method for evaluating the relevance of rating scales as criterion measures. While it is, of course, impossible to directly measure (a), approximate measures of some of the other components of the equation are available [for example, mean score as a measure of leniency, standard deviation of scores as a measure of central tendency, mean (*r* to *z* transformed) intercorrelation of trait scores as a measure of halo]. If Guilford's model is correct, it can be argued that the smaller the contribution of sources (b), (c), (d), and (e) to the given rating, the larger the contribution of (a), and therefore the greater the relevance of the given rating as a performance measure (assuming, of course, that the measure of trait J has indeed been assessed as content valid and is reliable).

Despite Conrad's (1932a, 1932b, 1933) contention that the problem of the "personal equation" in ratings is too small to be concerned about,

the error reduction approach described above has become a major method for evaluating the validity of rating scales and the utility of methods hypothesized to increase their validity. A small sample of the studies which have employed this approach includes those of Barrett et al. (1958), Bass (1956), Berkshire and Highland (1953), Borman and Dunnette (1975), Burnaska and Hollmann (1974), Campbell, Dunnette, Arvey, and Hellervik (1973), Creswell (1963), Keaveny and McGann (1975), Stockford and Bissell (1949), Taylor et al. (1958), Taylor and Hastman (1956), and Taylor and Wherry (1951).

The Analysis of Variance Approach

In addition to the error reduction approach described above, Guilford's (1954) work led to a second method for evaluating the validity of ratings--the analysis of variance approach. Originally based on Guilford's (1954, pp. 178-181) formulations, this method has been examined, refined and/or employed by Blumberg, DeSoto, and Kuethe (1966), Borman (1978), Boruch, Larkin, Wolins, and MacKinney (1970), Burnaska and Hollmann (1974), Friedman and Cornelius (1976), Johnson and Vidulich (1956), Kavanagh, MacKinney, and Wolins (1971), Stanley (1961), and Willingham and Jones (1958). When raters rate a number of ratees on a set of traits, seven sources of variance can be separated: ratees, traits, raters, ratees x traits interaction, ratees x raters interaction, traits x raters interaction, and ratees x traits x raters interaction. Blumberg et al. (1966) analyzed each of these seven sources of variance and found all but one of them to be related to the classical rating errors described above. The exception was the ratees x traits interaction

which "indexes the degree to which [ratees] are given distinct trait profiles on which [raters] agree. Since such differentiation is the goal of ratings, this is the component which is most likely to contain useful information" (p. 245). However, according to their analysis, "even this component may contain error, namely, stereotyping" (p. 245).

By obtaining a set of ratings and subjecting them to analysis of variance, it is possible to evaluate the contribution of each of the sources to the overall variance. If one includes as an additional experimental variable levels of some treatment hypothesized to affect these other sources of variance, then the treatment's effects can be assessed directly. Blumberg et al. (1966) advise: "If external measures of rating validity are unavailable, a reasonable objective would be to find a [treatment level] that maximizes [the ratees x traits] variance component" (p. 245). This approach was used successfully by Blumberg et al., Burnaska and Hollmann (1974), and Friedman and Cornelius (1976) to evaluate various rating scale formats in terms of validity.

Wherry's Theory of Rating

Wherry's (1952) comprehensive theory of rating, developed for the Department of the Army, takes the form of a mathematical equation based on the theoretical contributions of five writers. These contributions are Gulliksen's (1950) synthesis of the rationale of mental test theory and derivation of most of its theorems; Mosier's (1940) demonstration that both mental test theory and psychophysical theory stem from the same data--the interaction of persons and stimuli--and share the same basic theorems; Helson's (1947) "adaptation level" concept; Bellows' (1941)

specification of environmental contaminants to performance criteria; and Bartlett's (1932) theory of memory, which stresses the positive character of forgetting.

Wherry (1952) describes his equation as long and involved, quite complex, and "replete with unknown constants" (p. 7). "However," he argues, "it represents fairly well the actually complex response which rating involves" (p. 7). By manipulating the various terms in his formula, Wherry developed a system of theorems and corollaries to guide research. Not all of Wherry's theorems and corollaries have been tested empirically, and not all of those which have been tested have proven valid. Nonetheless, Wherry's theory is certainly one of the best developed systematizations of the rating process available, and has served well as a guide for research.

Graham's General Behavior Equation

The "general behavior equation" proposed by Graham (1950) can be used as a model to systematize the body of literature regarding the variables influencing the rating process. Graham's equation expresses an organism's response as a function of: (a) aspects of the stimulus; (b) the number of times the stimulus has been applied to the organism; (c) time; and (d) internal conditions of set, motivation, etc. In the typical psychophysics experiment the interest lies in studying response as a function of some stimulus property (a); thus it is necessary to hold other influencing variables (b), (c), (d) constant. According to Graham, studies involving the effect of (b) upon response are placed in the category of learning, those involving the relation of response to (c)

are in the category of forgetting and fatigue, and those relating response to (d) are in the area of motivation or emotion. The challenge to psychologists offered by Graham is to specify all of the variables in the general behavior equation, as well as the functional relationship which describes their influence on any given response.

As stated above, the judgment response in a psychophysics experiment should be influenced only by the particular stimulus characteristic under study. Given strict experimental controls in the laboratory setting, it is sometimes possible to restrict the situation appropriately. However, when the rating method is adapted for use in the field, rigid controls are often lacking. While the

$$R_{pac} = f(a) \qquad (1)$$

situation, where R_{pac} = performance rating of person p on behavior characteristic a by rater c, and

A = person p's true level of performance on behavior characteristic a

is what is desired in performance appraisal rating, the actual function is more likely to resemble the "general behavior equation" in the typical field application of the rating method. Viewed in terms of Graham's equation, the challenge facing industrial-organizational psychologists is to identify the other variables which influence R_{pac} in the applied situation, to specify the nature of the mathematical function, and to devise methods with which to control the other variables and reduce the

situation to a simple equation (1) relationship. This challenge is not likely to be met in the near future; however, industrial-organizational psychologists have made progress at least in the identification of some of the major categories of variables which must be included in any "general rating equation." Useful classifications of the variables influencing variance and error in ratings have been presented by Bass and Barrett (1972, pp. 228-238), Ghiselli and Brown (1955, pp. 88-91), Guilford (1954, pp. 320-321), Jenkins (1946), Lawler (1967), Lifson (1953), Nagle (1953), and Thorndike and Hagen (1969, pp. 424-431). The influence of their work is evident in the taxonomy described below.

A Taxonomy of Sources of Variance and Error in Ratings

This taxonomy, intended to systematize the literature regarding variables which influence rating judgments in the field, is organized in terms of a particular viewpoint of the rating process. According to this view, the judgmental process of performance rating takes place as follows: A rater uses a given instrument in a particular rating context and temporal situation to evaluate a ratee's performance on a given behavioral characteristic in a specific behavioral context. The outcome, a rating, may therefore be influenced by aspects of all of these categories of variables, and by interactions among them. Following Graham (1950), this viewpoint may be expressed as follows:

$$R = f(RTR_1 \ldots RTR_n,\ I_1 \ldots I_n,\ RC_1 \ldots RC_n,\ T_1 \ldots T_n,\ RTE_1 \ldots RTE_n,\ P_1 \ldots P_n,\ C_1 \ldots C_n,\ BC_1 \ldots BC_n,\ IA_1 \ldots IA_n,\ E) \qquad (2)$$

where R = obtained rating

$RTR_1 \ldots RTR_n$ = aspects of the rater

$I_1 \ldots I_n$ = aspects of the rating instrument

$RC_1 \ldots RC_n$ = aspects of the rating context

$T_1 \ldots T_n$ = aspects of the temporal situation

$RTE_1 \ldots RTE_n$ = aspects of the ratee

$P_1 \ldots P_n$ = aspects of the ratee's performance

$C_1 \ldots C_n$ = aspects of the behavioral characteristic

$BC_1 \ldots BC_n$ = aspects of the behavioral context

$IA_1 \ldots IA_n$ = interactions among various aspects of the other categories of variables

E = residual error.

Note that the nature of the mathematical function is not specified. It may take a simple linear form, such as the equation presented by Guilford (1954, p. 281), or a more complex linear combination of weighted (perhaps nonlinear) components of the type formulated by Wherry (1952, p. 7), or even a nonlinear logarithmic or power function of the type explored by Stevens (1960) and Whitlock (1963). To place this equation into perspective, recall that what is desired in performance rating is an equation of the form

$$R_1 = f(P_1). \tag{3}$$

The taxonomy of sources of variance and error in ratings is presented in outline form in Table 1. Note that Table 1 includes only

Table 1. A Taxonomy of Sources of Variance and Error in Ratings

I. Aspects of the Rater

A. Demographic Characteristics

1. Age and tenure (Mandell, 1956)
2. Sex (Hart & Olander, 1924)
3. Hierarchical position (Besco & Lawshe, 1959; Borman, 1974; Draper, 1964; Klimoski & London, 1974; Parker et al., 1959; Rambo, 1958; Springer, 1953; Tucker et al., 1967)

B. Ability Factors

1. Aptitude and achievement (Schneider & Bayroff, 1953)
2. Intelligence (Stockford & Bissell, 1949; Taft, 1955)
3. Effectiveness as a supervisor (Kirchner & Reisberg, 1962; Mandell, 1956; Levy & Stone, Note 4)

C. Motivational Factors

1. Cooperation and interest (Conrad, 1932b; Taft, 1955; Thorndike & Hagen, 1969, pp. 425-426)
2. Intentions (Bass, 1956; Cronbach, 1970, p. 577; Ronan, 1970)
3. Frustration (Rowland, 1970, p. 271; Smith & Kendall, 1963)

D. Personality Factors

1. "Yea-saying" versus "nay-saying" (Bass, 1956; Bass & Barrett, 1972, p. 233; Stockford & Bissell, 1949)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

2. Implicit trait theories (Bruner & Tagiuri, 1954; Cronbach, 1955; Gage & Cronbach, 1955; Hausman & Strupp, 1955)
3. Possession of desirable/undesirable traits (Hollingworth, 1922)
4. Self-image (Lundy, 1958; Vroom, 1959)

E. Knowledge Factors

1. Familiarity with the stimulus being rated (Blum & Naylor, 1968, p. 220; Christal & Madden, 1960; Madden, 1960, 1961)
2. Knowledge of appropriate norms and standards of performance (King, Ehrmann, & Johnson, 1952; Lifson, 1953; Taft, 1955)
3. Knowledge of common rating errors and how to avoid them (Borman, 1975; Driver, unpublished, described by McCormick & Tiffin, 1974, p. 215; Guilford, 1954, p. 295; Latham, Wexley, & Pursell, 1975)

II. Aspects of the Rating Instrument

A. General Formatting Factors (Barrett, 1966; Bass & Barrett, 1972, p. 229; Bayroff, Haggerty, & Rundquist, 1954; Blumberg et al., 1966; Dyer, Matthews, Stulac, Wright, & Yudowitch, 1975; Edwards, 1957; Freyd, 1923; Guilford, 1954, pp. 267-268; Madden & Bourdon, 1964; Stockford & Bissell, 1949; Taylor & Hastman, 1956; Uhrbrock, 1961)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

B. Anchoring Factors

1. Ambiguity/clarity of anchors (Bass & Barrett, 1972, p. 229; Champney, 1941; Cronbach, 1970, pp. 571-573; Guilford, 1954, p. 293)
2. Nature of the anchors--numerical/alphabetical vs. descriptive adjective vs. man-to-man vs. behavior-sample (Barrett et al., 1958; Borman & Dunnette, 1975; Borman & Vallon, 1974; Burnaska & Hollmann, 1974; Campbell et al., 1973; Dyer et al., 1975, p. VI-11; Ghiselli & Brown, 1955, pp. 104-108; Keaveny & McGann, 1975; Madden, 1964; Marsh & Perrin, 1925; Peters & McCormick, 1966; Ross, 1966; Smith & Kendall, 1963)
3. Number of anchors (Benjamin, 1952; Champney & Marshall, 1939; Dyer et al., 1975, pp. VI-1-9; Garner, 1960; Guilford, 1954, pp. 289-291; Symonds, 1924)
4. Balanced vs. unbalanced anchors (Weiss, 1963)

C. Focus Factors

1. Descriptive vs. evaluative focus (Stockford & Bissell, 1949)
2. Focus on past or present performance vs. future promise (Paterson, 1923; Smith & Kendall, 1963)

D. Forced Distribution Requirements (Klores, 1966)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

E. Standard of Comparison--Others vs. Absolute vs. Job (Barrett, 1966, pp. 79-85)

III. Aspects of the Rating Context

A. Opportunity to Observe (Ferguson, 1949a; Ghiselli & Brown, 1955, p. 90; Landy & Guion, 1970; Smith, 1976, p. 762; Thorndike & Hagen, 1969, pp. 427-428)

B. Purpose for Obtaining Ratings (Guilford, 1954, p. 295; Hollander, 1957; Taylor & Hastman, 1956; Taylor & Wherry, 1951)

C. Confidentiality (Bayroff et al., 1954; Creswell, 1963; Guilford, 1954, p. 295; Guion, 1965, p. 111; Paterson, 1923; Stockford & Bissell, 1949)

D. Organizational Factors

1. Organizational unit (McCormick & Tiffin, 1974, p. 210)
2. Support and concern shown by top management (Bass & Barrett, 1972, p. 235; Davis, 1953)
3. Organizational climate (Friedman & Cornelius, 1976, p. 215; Grey & Kipnis, 1976)

E. Training (Bernardin, 1978; Bernardin & Walter, 1977; Borman, 1975; Driver, unpublished, described by McCormick & Tiffin, 1974, p. 215; Guilford, 1954, p. 295; Kingsbury, 1922; Latham et al., 1975)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

F. Rater Participation in Scale Construction (Borman & Vallon, 1974; Campbell et al., 1973; Friedman & Cornelius, 1976; Smith & Kendall, 1963)

G. Presence of Supervision (Taylor & Hastman, 1956)

H. Number of Ratees Evaluated in One Session (Bayroff et al., 1954)

I. Presence of Environmental Stressors (Griffitt, 1970; Griffitt & Veitch, 1971; Sauser, Arauz, & Chambers, 1978; Arauz, Note 5)

IV. Aspects of the Temporal Situation

A. Time Available for Making the Ratings (Bayroff et al., 1954; Conrad, 1932b; Ghiselli & Brown, 1955, p. 91; Guilford, 1954, p. 294)

B. Temporal Variables Influencing Judgment (Guilford, 1954, pp. 302-311)

C. Time Interval Under Consideration (Bernardin, 1978; Ghiselli & Brown, 1955, p. 81)

D. Acclimation to the Job Over Time (Bass, 1962; Fleishman & Fruchter, 1960; Fleishman & Hempel, 1954; Ghiselli & Haire, 1960; Hollander, 1957)

V. Aspects of the Ratee

A. Race (Farr, O'Leary, & Bartlett, 1971; Greenhaus & Gavin, 1972; Schmidt & Johnson, 1973)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

B. Sex (Deaux & Emswiller, 1974; Goldberg, 1968; Jacobson & Effertz, 1974; Pheterson, Kiesler, & Goldberg, 1971; Rosen & Jerdee, 1973, 1974)

C. Age and Tenure in Present Position (Bass & Barrett, 1972, pp. 231-232; Rothe, 1949)

D. Job Level (Klores, 1966; Levine & Butler, 1952; McCormick & Tiffin, 1974, p. 210)

E. Reputation (Hemphill & Sechrest, 1952)

VI. Aspects of the Ratee's Performance

A. Correctness (Gordon, 1970)

B. Variability (Ayers, 1942; Carter & Dudek, 1947; Hay, 1943; Jenkins, 1946; Klemmer & Lockhead, 1962; MacKinney & Wolins, 1960; Owens, 1942; Ronan & Prien, 1966; Rothe, 1946a, 1946b, 1947, 1951, 1978; Rothe & Nye, 1958, 1959, 1961; Scott & Hamner, 1975; Seashore, 1931; Ronan, Note 7)

VII. Aspects of the Behavioral Characteristic

A. Overtness (Ferguson, 1949a; Paterson, 1923; Stockford & Bissell, 1949; Thorndike & Hagen, 1969, pp. 428-429)

B. Dimensionality (Cronbach, 1970, p. 572)

VIII. Aspects of the Behavioral Context

A. Physical Working Conditions (Bellows, 1941; Brogden & Taylor, 1950b; Ronan & Prien, 1966; Toops, 1944; Ronan, Note 1)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

B. Units of Production (Toops, 1944)

C. Work Group Rate Restrictions (Lupton, 1963; Viteles, 1932, pp. 560-565, 1953, pp. 45-61; Whyte, 1955)

D. Family and Ethnic Customs (Viteles, 1932, pp. 213-214)

E. Attitudes, Perceptions, and Opinions (Patchen, 1970; Viteles, 1936)

IX. Interactions

A. Rater x Ratee Interactions

1. Race of rater x race of ratee (Cox & Krumboltz, 1958; deJung & Kaplan, 1962; Flaugher, Campbell, & Pike, 1969)

2. How long rater has known ratee (Ferguson, 1949a; Stockford & Bissell, 1949)

3. Friendship between rater and ratee (Hollander, 1956)

4. Rater's perceived similarity to ratee (Bass & Barrett, 1972, p. 234; Lundy, 1958; Vroom, 1959)

5. Rater's perception of ratee's similarity to members of rater's family (Campbell & Chapman, 1957)

B. Ratee x Ratee's Performance Interactions

1. Sex of ratee x ratee's performance (Bigoness, 1976; Deaux & Taynor, 1973)

2. Race of ratee x ratee's performance (Bigoness, 1976)

Table 1. A Taxonomy of Sources of Variance and Error in Ratings (Cont'd)

C. Rater's Personality x Ratee's Performance (Wexley et al., 1972)

D. Sex of Ratee x Overtness of Behavioral Characteristic (Deaux & Emswiller, 1974; Pheterson et al., 1971)

E. Sex of Ratee x Ratee's Hierarchical Position (Jacobson & Effertz, 1974)

F. Rater's Cognitive Complexity x Scale Format (Schneier, 1977)

G. Interactions with Training

1. Rater's intelligence x training (Stockford & Bissell, 1949)
2. Race of ratee x training (Schmidt & Johnson, 1973)
3. Format of scale x training (Brown, 1968)
4. Time since training x training (Bernardin, 1978)

H. Higher Order Interactions

1. Race of rater x race of ratee x ratee's performance (Hamner, Kim, Baird, & Bigoness, 1974; Rotter & Rotter, Note 8)
2. Rater's implicit personality theory x how long rater has known ratee (Koltuv, 1962)

Note. The table includes only those sources which have been documented in the literature; other sources may exist. Citations listed after each source are representative of theoretical or empirical evidence for the source.

those sources of variance and error which have been documented in the literature. Many other potential entries exist--the cataloging of interactions among sources in the other categories, for example, has only reached the tip of the iceberg of potential. As Friedman and Cornelius (1976) point out, a systematic analysis of the various factors that influence ratings in applied settings, especially the interactive factors, is "sorely needed" (p. 216). Such an analysis should seek to determine how the various sources of variance relate to each other as well as to characteristics of ratings. It is quite possible that some of the variables listed in Table 1 mediate the effects of other listed variables. Only when the complex interplay among the variables in Table 1 and the diverse characteristics of ratings has been systematically examined through a program of controlled experimentation will the true nature of the rating process be revealed.

Approaches to Reducing Error in Ratings

A Variety of Approaches

The identification of sources of variance and error in ratings has enabled researchers to develop several relatively successful approaches to ridding the rating process of error, thus theoretically enhancing the validity of ratings. One such approach, focused on the elimination of unwanted variance due to factors listed primarily under categories II and VII of Table 1, is to restructure the focus and format of the rating scale. Examples of applications of this approach include the use of forced distribution requirements (Klores, 1966), the scientific determination of the proper number and placement of anchors

(Champney & Marshall, 1939; Garner, 1960), the shift in emphasis from the evaluation of vague personality traits to the description of overt, observable behaviors (Campion, 1972; Flanagan, 1954; Ghiselli & Brown, 1955, pp. 104-108; Smith & Kendall, 1963; Stockford & Bissell, 1949), and the replacement of numerical and alphabetical anchors with more meaningful anchoring cues, such as descriptive adjectives and even samples of behavior (Campion, 1972; Flanagan, 1954; Ghiselli & Brown, 1955, pp. 104-108; Peters & McCormick, 1966; Smith & Kendall, 1963).

A second approach to reducing the amount of error in ratings of performance, focused primarily on the contributions of variables listed in categories III and IV, is to restructure the rating context. Specific examples of this approach include emphasizing managerial concern and support for the rating process (Davis, 1953), enhancing the opportunity to observe ratee performance and structuring the observation process (Bernardin & Walter, 1977; Burke & Goodale, 1973; Flanagan & Burns, 1955; Ghiselli & Brown, 1955, p. 90; Tate, 1964; Zedeck & Baker, 1972), attending to the issues of purpose and confidentiality in obtaining ratings (Bayroff et al., 1954; Stockford & Bissell, 1949; Taylor & Hastman, 1956; Taylor & Wherry, 1951), and providing more time for raters to do an adequate job of evaluating performance (Bayroff et al., 1954; Conrad, 1932b).

A third approach to error reduction has concentrated on eliminating category VIII variance by standardizing the context in which the ratees perform. This approach is exemplified by several techniques discussed earlier--simulation (Besnard & Briggs, 1967; Viteles, 1945),

assessment centers (Bray, 1964; Bray & Campbell, 1968; Bray & Grant, 1966; Byham, 1970; Byham & Thornton, 1970; Finkle, 1976), and the "psychometric approach to job performance" (Atlanta Regional Commission, 1974; Ronan et al., 1976; Talbert et al., 1976).

A fourth category, focused on eliminating variance due to aspects of the rater (category I) and the ratee (category V), as well as interactions among them (category IX), has been suggested but rarely applied. This approach consists of identifying "good" raters and "rateable" ratees and structuring the rating situation such that only "rateable" ratees are evaluated, and then only by "good" raters. Prototypes of this approach have been suggested by Cronbach (1970, p. 577), Mullins and Force (1962), and Wiley and Jenkins (1964). Borman (1974) has offered a somewhat similar suggestion: Use multiple raters, each focusing on those aspects of the ratee's performance with which he is most familiar. The rarity of application of this approach is most likely due to its impracticality in the applied setting--availability of multiple raters is rare, and organizational requirements typically demand that all ratees be appraised on some predesignated set of variables.

Two additional approaches, rater training and rater participation in scale construction, have also been found to be effective in reducing the contributions of extraneous variables listed in Table 1. Since these two approaches are examined in the present study, they are discussed in more detail below.

Rater Training

Importance of Training. "Various experiences with ratings tend to show that the most effective method for improving ratings in many ways is to train raters carefully" (Guilford, 1954, p. 280). This statement has been echoed by numerous writers in the field of performance appraisal (Barrett, 1966, p. 120; Baylie et al., 1974, p. 165; Bittner, 1948, p. 419; Kingsbury, 1922, pp. 377-378; McCormick & Tiffin, 1974, p. 215; Smith, 1976, p. 762; Thorndike & Hagen, 1969, p. 446). Despite the apparent importance of rater training in error reduction, such training programs are rarely offered or emphasized in industrial organizations (Bittner, 1948, p. 420; Spicer, 1951; Lopez, Note 8). While rater training is not always effective in reducing error (Katzell, unpublished, described by Barrett, 1966, p. 125; Taylor & Hastman, 1956), the weight of empirical evidence favors training for raters, "even though some kinds of training, under some circumstances, do not "help" (Barrett, 1966, p. 125).

Empirical Evidence. Bittner (1948, pp. 421-422) describes two controlled studies of the effects of rater training conducted by the Army Personnel Research Section during World War II. The first experiment involved 603 officers rating 2401 men:

> One group was given training consisting of a two-hour period of instruction on the basic principles of accurate rating, the meaning of the rating scale traits and the numerical points on the scales, how to use the rating form, a practical problem in rating, and a test over what they had been taught. The other group was given no training. (p. 421)

The results: Training increased the accuracy of the ratings and decreased the effects of leniency error. The second of Bittner's studies

revealed that the inclusion of training materials in the rating form also reduced error.

Similar findings characterize the results of additional rater training studies. Stockford and Bissell (1949) found that a six-hour rater training program was significantly more effective than a two-hour general orientation session in reducing bias on an industrial rating scale. Bayroff and Burke (1950) developed a 24-page "rater's guide" which they claimed was useful in helping enlisted army personnel reduce the extent of certain errors in performance evaluation ratings. King, Erhmann, and Johnson (1952) found significant increases in the inter-rater reliability of judgments of children's social behavior when the raters "examined each other's ratings and attempted to establish common criteria for each item" (pp. 152-153). Levine and Butler (1952) found that a group discussion of error in ratings significantly decreased bias due to ratee's job level; however, training via formal lecture had no such effect. Ryder (1962) developed a flashcard training system which he found effective in increasing interrater reliability of clinical ratings of behavior. Brown (1968) found that a rater training program similar to that described by Bittner (1948) significantly reduced halo error in peer ratings of student nurses.

Recent studies of rater training programs have continued to document their successful application. Schmidt and Johnson (1973) presented evidence suggesting that training raters in human relations may eliminate race effects in rating. Wexley, Sanders, and Yukl (1973) found an intensive training workshop to be successful in eliminating

contrast effects in employment interviews--after the use of warnings and special scale anchors failed to eliminate contrast error. Driver (unpublished, described by McCormick & Tiffin, 1974, p. 215) reported that a seven-hour training program in methods of rating succeeded in reducing halo effect. Latham, Wexley, and Pursell (1975) found both group discussion and lecture/workshop training methods to be effective in reducing similarity, contrast, and halo errors. Contrary to Levine and Butler's (1952) conclusions, Latham et al. (1975) found the lecture/workshop method to be superior to the group discussion method both in reducing "first impression" errors and in popularity with the trainees. Borman (1975) reported that a five-minute training session "significantly reduced halo, while leaving validity of the ratings generally unaffected" (p. 556). However, states Borman, "Performance ratings completed after training possessed lower reliability, although raters provided somewhat more accurate performance profiles" (p. 556). Bernardin and Walter (1977) found one hour of training to significantly reduce halo error and, when combined with full exposure to the rating scales during training, to reduce leniency error and increase interrater reliability as well. Bernardin (1978) found that a comprehensive rater training program was more effective in reducing leniency and halo errors than were an abbreviated program or no program. Unfortunately, the effects of training were found to diminish rapidly over time. In his most recent paper, Bernardin again reports evidence supporting the effectiveness of a rater training program (Bernardin & Boetcher, Note 9).

Training Program Content. Regarding the content of a rater training program, Bittner (1948, pp. 425-426) prescribed the following:

1. Clarification of the aims and purposes of merit rating.
2. Instruction on the meaning of characteristics or traits to be evaluated.
3. Instruction on the meaning of the points on the scale.
4. Instruction on the avoidance of common pitfalls in rating such as:
    a. Lack of objectivity--basing ratings on supposition, guesswork, emotional bias.
    b. Rating one trait in the light of ratings on other traits.
    c. Rating on the basis of general impressions.
    d. Rating on the basis of a single dramatic incident.
    e. Restricting the spread of ratings.
5. Supervised practice and discussion of practice ratings made.
6. Instruction in how to use and interpret the ratings.
7. Periodic refresher training.

Empirical evidence documents the effects of three major components of the training program (Brown, 1968). These are: (a) practice with the specific scales to be used in the rating program (Bernardin & Walter, 1977; Wakeley, Note 10); (b) discussion of errors in rating by the raters (Latham et al., 1975; Levine & Butler, 1952), and (c) special emphasis on the importance of trait differentiation (Latham et al., 1975; Taylor & Hastman, 1956).

Mediating Variables. The evidence described above strongly suggests that rater training is an effective method for reduction of error variance in ratings. The evidence regarding the processes whereby rater training programs work is not nearly so clear; however, research and theory suggests that two major variables appear to mediate the effects of training on characteristics of ratings. These are: (a) attitudes toward the rating process and motivations toward rating (Barrett, 1966, p. 121: Bittner, 1948, pp. 422-424; Brown, 1968; Levine & Butler, 1952; Ryder, 1962; Thorndike & Hagen, 1969, p. 446); and (b) knowledge of the performance appraisal rating process, including the identification and avoidance of common rating errors (Bernardin, 1978; Bernardin & Walter, 1977; Bittner, 1948, p. 424; Borman, 1975; Guilford, 1954, p. 295; Latham et al., 1975; Levine & Butler, 1952; Thorndike & Hagen, 1969, p. 446; Wexley et al., 1973). Thus, according to what is known about rater training, the major reasons it appears to work are that it (a) enhances motivation and attitude toward rating and (b) sensitizes raters to the rating process and provides raters with the knowledge of how to identify and avoid errors. These two mediating variables, in turn, appear to affect such characteristics of ratings as reliability, constant errors, and discriminant and convergent validity.

Rater Participation in Scale Construction

Importance of Participation. "Research has shown that impressive increases in productivity can be brought about by giving employees a greater opportunity to participate in decision making" (Vroom, 1976, p. 1538; see also Coch & French, 1948; Hunt, 1974; Lowin, 1968; Marrow,

Bowers, & Seashore, 1967; Morse & Reimer, 1956; Tannenbaum, 1966, pp. 84-102; Vroom, 1969; Wood, 1973). While participation is not an organizational panacea (Fleishman, 1965; French, Israel, & As, 1960), its usefulness as a tool for successfully accomplishing organizational change has led several writers to suggest its incorporation into the performance appraisal rating process (Barrett, 1966, p. 14; Baylie et al., 1974, p. 170; Bittner, 1948, p. 423; French, Kay, & Meyer, 1966; Friedman & Cornelius, 1976; Meyer, Kay, & French, 1964; Rundquist & Bittner, 1950; Smith, 1976, p. 762; Smith & Kendall, 1963). Friedman and Cornelius (1976, p. 215) have summarized several speculations on why rater participation may enhance positive characteristics of rating scales. These include: (a) increased understanding of the job being rated and its various components (Smith & Kendall, 1963); (b) positive effects of raters' expectancies, valences, or instrumentalities with respect to scale use (Mitchell, 1974; Vroom, 1964); (c) increased effort or conscientiousness when rating due to cognitive dissonance (Festinger, 1957); and (d) greater acceptance and commitment to scales resulting from group decisions (Lowin, 1968; Maier, 1967; Vroom & Yetton, 1973; Wood, 1973). Three studies demonstrating the effectiveness of the use of participative techniques in rater training programs have already been presented (King et al., 1952; Latham et al., 1975; Levine & Butler, 1952). The majority of the evidence regarding the effectiveness of rater participation in scale construction, however, centers around behaviorally anchored rating scales (BARS) and the retranslation technique (Smith & Kendall, 1963) described below.

Behaviorally Anchored Rating Scales and the Retranslation Technique. Drawing on the heritage of research with the critical incident technique (Flanagan, 1949, 1954; Flanagan & Burns, 1955), the use of behavioral descriptors as scale anchors (Barrett et al., 1958; Ghiselli & Brown, 1955, pp. 107-108; Peters & McCormick, 1966; Schultz & Siegel, 1961), the benefits of employee participation (Bittner, 1948, p. 423; Coch & French, 1948; Rundquist & Bittner, 1950), and language translation methodology, as well as their own ingenuity, Smith and Kendall (1963) developed a participative technique for constructing behaviorally anchored rating scales (BARS) of employee performance--the retranslation technique. The philosophical underpinnings of this technique were: (a) do not trick the rater--help him; (b) use words, dimensions, and anchors that are relevant to the rater; and (c) involve the rater in the scale development process (Smith & Kendall, 1963).

While numerous variations on the BARS development technique have evolved (Arvey & Hoyle, 1974; Bernardin, LaShells, Smith, & Alvares, 1976; Campbell et al., 1973; Dickinson & Tice, 1973; Hoyle & Arvey, 1972; Kafry, Zedeck, & Jacobs, 1976; Schwab, Heneman, & DeCotiis, 1975; Smith & Kendall, 1963; Tate, 1964; Zedeck, Kafry, & Jacobs, 1976), the general procedure involves seven steps: (a) identify dimensions of performance, (b) gather critical incidents illustrating each dimension, (c) have participants independently classify incidents, (d) eliminate unclear incidents and dimensions, (e) have participants independently evaluate the remaining incidents on a scale of desirability, (f) eliminate unclear incidents, and (g) construct behavioral expectation scales anchored by scaled incidents.

Since the development of the retranslation technique, BARS have been constructed to evaluate the performance of nurses (Burke & Goodale, 1973; Smith & Kendall, 1963; Tate, 1964; Zedeck & Baker, 1972; Zedeck et al., 1974), college professors (Bernardin, Alvares, & Cranny, 1976; Bernardin, LaShells, Smith, & Alvares, 1976; Burnaska & Hollmann, 1974; Friedman & Cornelius, 1976; Harari & Zedeck, 1973; Keaveny & McGann, 1975; Zedeck, Jacobs, & Kafry, 1976), department store managers (Campbell et al., 1973), grocery clerks (Fogli, Hulin, & Blood, 1971), electronic data processing specialists (Arvey & Hoyle, 1974; Hoyle & Arvey, 1972), naval officers (Borman & Dunnette, 1975), engineers (Williams & Seiler, 1973), police officers (Cascio & Valenzi, 1977; Landy, Farr, Saal, & Freytag, 1976), secretaries (Borman, 1974), and a cluster of diverse hospital jobs (Goodale & Burke, 1975). "Spin-offs" include suggestions for training programs (Blood, 1974) as well as scales for measuring motivation (Landy & Guion, 1970), morale (Motowidlo & Borman, 1977), and interviewee qualifications (Maas, 1965).

Research with BARS has shown that while they fall short of being the ideal performance appraisal technique, they do appear to have some value for use in the applied setting (Campbell et al., 1970, pp. 118-125; Dunnette, 1966, pp. 95-100; Schwab et al., 1975). Reliability estimates for ratings with BARS are medium to high (Burnaska & Hollmann, 1974; Fogli et al., 1971; Landy et al., 1976; Smith & Kendall, 1963), and BARS seem to possess adequate convergent validity, but a number of studies have questioned their discriminant validity (Arvey & Hoyle, 1974; Campbell et al., 1973; Dickinson & Tice, 1973, 1977; Friedman &

Cornelius, 1976; Keaveny & McGann, 1975; Williams & Seiler, 1973; Zedeck & Baker, 1972). One study (Zedeck & Baker, 1972) presented evidence of a weak relationship between rated performance and an objective measure of performance (tenure). However, Cascio and Valenzi (1978) found much stronger relationships among BARS and objective measures purported to measure the same dimensions of police officer performance. BARS are typically found to be superior to numerical/alphabetical- and adjective-anchored rating scales in terms of error reduction (Borman & Dunnette, 1975; Burnaska & Hollmann, 1974; Campbell et al., 1973; Keaveny & McGann, 1975). However, this is not always the case, as Bernardin (1977), Bernardin, Alvares, and Cranny (1976), Borman and Vallon (1974), and Friedman and Cornelius (1976) have shown. This latter set of studies suggests that it is the process of involving raters in scale construction, not the characteristics of the resulting scales, which leads to error reduction.

Participation as the Key to Success. While Smith and Kendall (1963) apparently intended to involve all potential raters in the BARS development technique, thus taking advantage of the effects of participation (Smith, 1976, p. 762), many of the studies reported above investigated BARS developed by one set of raters, but used by another. The contention that rater participation, rather than simply scale format, is the key to the success of the retranslation technique and its resulting BARS (Campbell et al., 1973; Smith & Kendall, 1963) has been given strong support by the findings of three recent studies. Borman and Vallon (1974) compared the BARS originally developed by Smith and

Kendall (1963) to appraise the performance of nurses with a set of numerically-anchored rating scales. The raters had no previous experience with either set of scales. In terms of interrater reliability and confidence in ratings, Borman and Vallon found the BARS superior. "However, when the simpler scale was used there was significantly less leniency effect and raters were better able to discriminate among different ratees in terms of performance" (p. 197). Borman and Vallon concluded:

> When a behavioral expectation scale is transported from one setting to another, the effectiveness of the scaled-expectations format may suffer because the raters do not participate in scale development and/or certain anchors are inappropriate for the new situation. (p. 197)

Bernardin, Alvares, and Cranny (1976) provide further support for the belief that rater participation is the key to successful rating scale development. Bernardin et al. took issue with Campbell et al.'s (1973) finding that BARS produce less lenient ratings than do summated rating scales (SRS). They claimed that the Campbell et al. methodology was biased in favor of the BARS, since ratees participated in developing the BARS, but not the SRS. In their own study, Bernardin et al. compared three sets of scales for leniency effects: (a) BARS developed by ratees using the retranslation technique, (b) SRS developed by the experimenters, and (c) SRS developed by the experimenters but involving the participation of raters in an item analysis procedure. While the BARS were superior to the first set of SRS in terms of leniency reduction, the item-analyzed SRS were found to be the most effective scales on the basis of this criterion.

Strong support for the importance of participation in scale construction comes from Friedman and Cornelius (1976), who compared sources of variance in ratings made by subjects in three conditions: (a) participation in developing a set of BARS; (b) participation in developing a set of graphic rating scales; and (c) no participation in scale development. They discuss their findings as follows:

> The results of the present study suggest that rater participation in scale construction led to greater convergent validity, less relative halo, as well as lower levels of variance attributable to rating errors. Participation did not lead to high levels of discriminant validity. Of particular importance is the finding that participation led to more desirable psychometric operating characteristics using either scale format. That is, regardless of which scale format was used, subjects who had participated in scale development provided ratings that were psychometrically superior. This phenomenon is consistent with Smith and Kendall's (1963) original, previously untested proposition that participation in scale development leads to increases in the validity of ratings. (p. 215)

Mediating Variables. The evidence described above suggests that rater participation in scale construction, like rater training, is an effective method whereby to reduce error variance in ratings. Again, as with rater training, the processes which mediate the effects of participation on characteristics of outcomes are not clearly understood, although Friedman and Cornelius (1976, p. 215) have made some valuable suggestions. Existing research and theory regarding rater participation does point to one major mediating variable: attitudes toward the rating process and motivations toward rating (Barrett, 1966, p. 14; Bittner, 1948, p. 423; Coch & French, 1948; Friedman & Cornelius, 1976; Levine & Butler, 1952; Rundquist & Bittner, 1950; Smith & Kendall, 1963). Thus, rater participation in scale construction appears to work because

it does one of the same things rater training does: it enhances motivation and attitude toward rating. As stated above, this mediating variable, in turn, appears to affect such characteristics of ratings as reliability, presence of error, and discriminant and convergent validity.

Comparisons of Training and Participation

From the evidence discussed above, it appears that both rater training and rater participation in scale construction are effective approaches to reducing the contributions of unwanted sources of variance in ratings. Several major questions remain unanswered, however. Three important questions which stimulated this dissertation research project are:

1. What are the relative effects of training and participation on characteristics of ratings? Some researchers argue that rater training is the essential process in obtaining valid ratings (Bernardin & Walter, 1977; Borman, 1975; Dickinson & Tice, 1973; Guilford, 1954, p. 280; Zedeck & Baker, 1972), some maintain that rater participation is the key (Borman & Vallon, 1974; Campbell et al., 1973; Friedman & Cornelius, 1976), still others claim that both processes are vital (Bittner, 1948; Smith, 1976, p. 762). A fourth viewpoint is that neither process is necessary: Zedeck et al. (1976) report that their subjects neither participated in scale development nor received training in their use, yet were able to use the scales effectively. Which approach, training or participation, is more effective in reducing sources of error? Do the two approaches differentially affect various characteristics of ratings? Are the two approaches interchangeable, complementary, unnecessary?

2. What are the relative effects of training and participation on mediating variables? Both training and participation are hypothesized to affect one mediating variable: attitudes toward the rating process and motivations toward rating. An additional mediator, knowledge of the rating process, including identification and methods for the avoidance of common rating errors, is hypothesized to be affected by training, and may be reached indirectly by participation. Which approach, training or participation, has stronger effects on these mediators? Do the two approaches differentially affect the mediators? Are the two approaches interchangeable, complementary, unnecessary?

3. How well do the hypothesized mediating variables explain the effects of the two approaches? If the variance due to the mediating variables is partialled out of the ratings, will the effects of the two approaches "wash out," or will there still be effects on certain characteristics of the ratings which must be explained above and beyond the variance accounted for by the mediators?

CHAPTER II

STATEMENT OF THE PROBLEM

A Hypothetical Model of the Effects of Training and Participation

A hypothetical model of the effects of rater training and rater participation in scale construction on characteristics of ratings is presented in Figure 1. This model is based on the literature reviewed in the previous chapter, and is intended to serve as a schematic guide to the three groups of questions posed at the end of Chapter I: (a) What are the relative effects of training and participation on characteristics of ratings? (b) What are the relative effects of training and participation on mediating variables? (c) How well do the hypothesized mediating variables explain the effects of the two approaches?

The model suggests that the effects of participation on characteristics of ratings are mediated by subjects' attitudes toward performance appraisal rating, while an additional mediating variable, subjects' knowledge of the performance appraisal rating process, is required to account for the effects of training. The strengths of the relationships among the variables are unspecified--this study was intended in part to determine those strengths. The particular characteristics of ratings hypothesized to be affected are also left unspecified in the diagram--a variety of characteristics were examined in this study.

The model is, of course, incomplete--it should be viewed as a subsystem of a larger model rather than as a closed-system entity. The

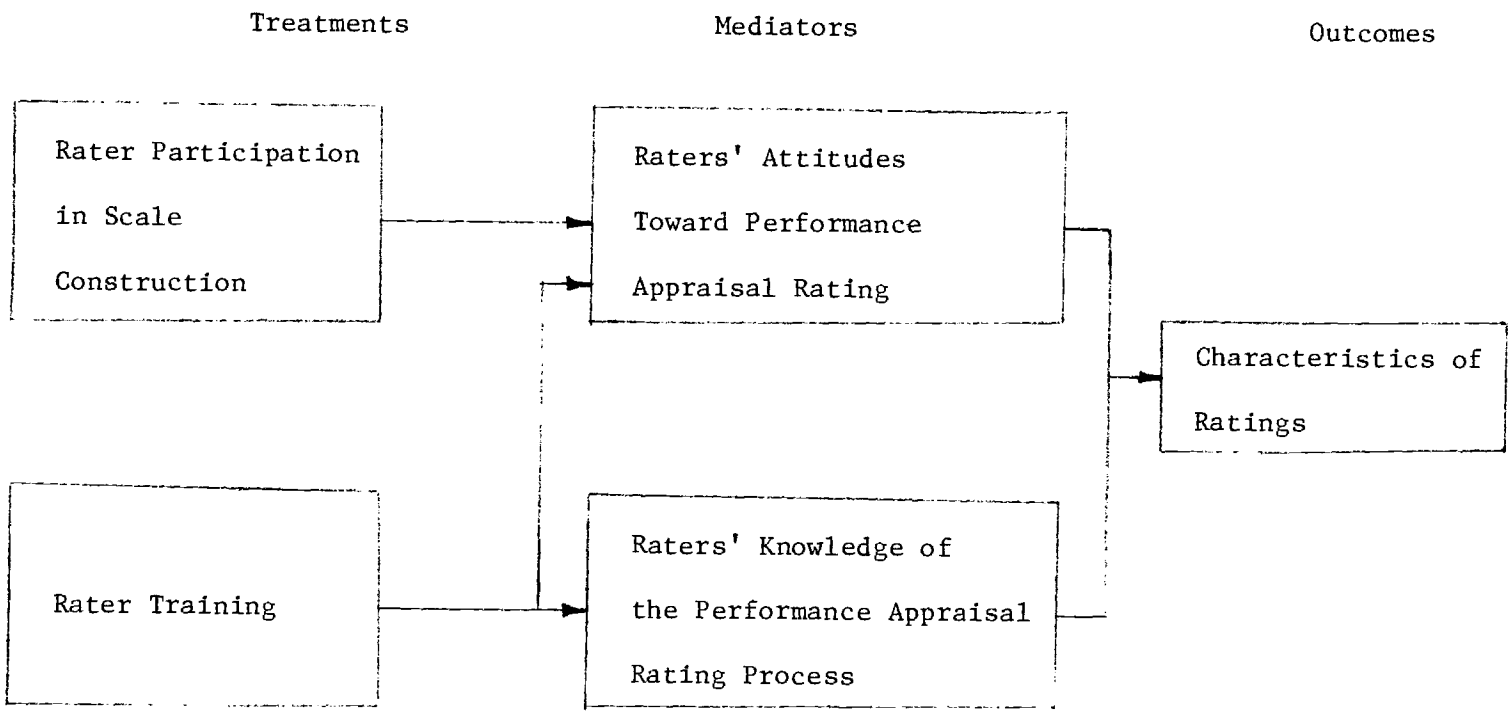


Figure 1. A Hypothetical Model of the Effects of Training and Participation

complete model would include all of the sources of variance and error presented in Table 1. The model diagrammed in Figure 1 expresses the hypothesized relationships among only those variables chosen for examination in this study.

Variables Under Consideration

Treatment Variables

Two "treatment" variables were considered in this investigation: (a) rater participation in scale construction, and (b) rater training. Operational definitions of these two variables are presented in the "Treatment Conditions and Procedures" section of Chapter III. The two treatments served as independent variables in all three phases of the data analysis.

Mediator Variables

As diagrammed in Figure 1, two "mediator" variables were included in this investigation: (a) raters' attitudes toward performance appraisal rating, and (b) raters' knowledge of the performance appraisal process. The mediator variables are operationally defined in the "Instruments and Materials" section of Chapter III. They served as dependent variables in the second phase of the analysis and as covariate variables in the third phase.

Outcome Variables

As described in detail in Chapter III, experimental subjects were asked to evaluate five simulated college professors on five categories of performance using a set of behaviorally anchored rating scales. Various characteristics of the resulting ratings were examined for effects

of training and participation in the first and third phases of the research project. These characteristics include sources of variance in the ratings, elevation and dispersion of the ratings, intercorrelation of the ratings, and intraclass and one-rater reliabilities and validities of the ratings. These outcome variables are operationally defined below and in Chapter IV, which presents the results of the ten analytic studies outlined below.

Plan of Analysis and Hypotheses

The present investigation included ten studies of the variables implied in the hypothetical model given in Figure 1. They were arranged in three phases, one corresponding to each of the three questions mentioned at the end of Chapter I.

Phase One: Effects of Treatments on Outcomes

The model indicates that rater participation in scale construction and rater training should have measurable effects on various characteristics of ratings. The five studies included in the first phase of the data analysis were intended to investigate this aspect of the model. They were designed not only to provide data to support inferences about the main effects of the two treatments, but also to examine possible interactions between the treatments. This information could be useful as a basis for conclusions regarding the use of the two treatments, or some combination of them, to affect various characteristics of ratings collected in the applied setting.

Study One. The purpose of this study was to examine the effects of Participation and Training on sources of variance in the obtained

ratings. The analysis employed was a modification of Guilford's (1954, pp. 278-281) analysis of variance approach, mentioned in Chapter I, as espoused by Blumberg et al. (1966), Borman (1978), Boruch et al. (1970), Burnaska and Hollmann (1974), Friedman and Cornelius (1976), Johnson and Vidulich (1956), Kavanagh et al. (1971), Stanley (1961), and Willingham and Jones (1958).

Sources of variance examined in this study included: (a) Treatments (Participation and Training), the overall elevation of the ratings made under each treatment, typically employed as a measure of the leniency error; (b) Ratees, effects of the individual ratees on the ratings, typically employed as a measure of consensual halo error; (c) Categories, effects of categories on the ratings, sometimes viewed as a measure of consensual contrast or comparison error; and (d) Ratees x Categories interaction, typically interpreted as discriminant validity. Higher-order interaction effects were also examined, allowing the estimation of the effects of the treatments on each of these sources of variance.

The general hypotheses[1] examined in this study were:

H_1 Training, Participation, and Training x Participation effects on overall elevation of the ratings are statistically

---

[1]The "general hypotheses" presented here and on the following pages actually represent sets of related statistical hypotheses. For example, general hypothesis H_1 includes the following specific statistical null hypotheses:

$H_{0_{1a}}$ Training has no effect on overall elevation of the ratings.

$H_{0_{1b}}$ Participation has no effect on overall elevation of the ratings.

$H_{0_{1c}}$ Training x Participation has no effect on overall elevation of the ratings.

significant. (Training and Participation are expected to result in significantly lower elevations.)

H_2 Training x Ratees, Participation x Ratees, and Training x Participation x Ratees effects are statistically significant. (Training and Participation are expected to result in significantly smaller Ratees effects.)

H_3 Training x Categories, Participation x Categories, and Training x Participation x Categories effects are statistically significant. (Training and Participation are expected to result in significantly smaller Categories effects.)

H_4 Training x Ratees x Categories, Participation x Ratees x Categories, and Training x Participation x Ratees x Categories effects are statistically significant. (Training and Participation are expected to result in significantly greater Ratees x Categories effects.)

Thus, Study One was intended to determine whether Participation and/or Training actually reduced elevation of ratings and unwanted variance attributable to Ratees and/or Categories, while increasing variance attributable to the desirable Ratees x Categories interaction. While statistically significant interaction effects between Participation and Training were predicted, there was not enough evidence available to specify the exact nature of the interactions. The actual analysis employed was a split-plot factorial ANOVA (see Kirk, 1968, pp. 311-312).

Study Two. The purpose of this study was to examine the effects of Training, Participation, and the Training x Participation interaction on elevation of ratings per category. As stated above, elevation is often used as an operational definition of the leniency error. The general hypothesis examined in this study was:

H_5 For each category, Training and Participation have significant main and interactive effects on elevation of the ratings. (Training and Participation are expected to result in significantly lower levels of elevation.)

The analysis employed in this study was a MANOVA, with each rater's mean rating (across ratees) for each of five categories serving as the dependent variables. Individual ANOVAs for each category were examined for interpretive purposes. Again, no predictions could be made regarding the exact nature of the anticipated significant interaction effect.

Study Three. This study was intended to examine the effects of Training, Participation, and the Training x Participation interaction on the dispersion of ratings per category. Dispersion (variance) is one measure which has been used to operationally define the central tendency error. The general hypothesis under examination in this study was:

> H_6 For each category, Training and Participation have significant main and interactive effects on dispersion of the ratings. (Training and Participation are expected to result in a significantly greater level of dispersion.)

This study also utilized a MANOVA design, with each rater's variance of ratings (across ratees) for each of five categories serving as the dependent variables. Individual ANOVAs for each category were examined for interpretive purposes. The nature of the expected significant interaction effect was once more left unspecified.

Study Four. The purpose of this study was to examine the effects of Training, Participation, and their combination on mean intercorrelation among category ratings, a statistic typically employed to operationally define the halo error. The general hypothesis examined in this study was:

> H_7 For each simulated ratee, Training, Participation, and their combination significantly affect mean r-to-Z transformed intercorrelation among category ratings. (Training and Participation are expected to result in significantly smaller mean r-to-Z transformed intercorrelations among categories.)

For each cell of the design, intercorrelations of category ratings (across subjects) were calculated for each simulated ratee and were converted to Z scores through the use of Pearson's r-to-Z transformation (Guilford & Fruchter, 1978, p. 522). The hypothesis was tested for each ratee by comparing the mean Z scores for the four cells for equivalence using Cohen and Cohen's (1975, p. 52) χ^2 test for homogeneity among independent sample correlation coefficients.

Study Five. The purpose of this study was to examine the effects of Training, Participation, and the Training x Participation interaction on the reliability and validity of the obtained ratings. Using Ebel's (1951; Guilford, 1954, pp. 395-397) formulae, intraclass reliability coefficients were calculated for each of five categories in each of the four cells of the Participation x Training design. Since, as discussed in the next chapter, the simulated ratees were constructed such that their "true scores" on each category were known, it was also possible to calculate validity coefficients for ratings of each category within each cell of the design. Two such coefficients were calculated: (1) the intraclass correlation coefficient between "true scores" and the mean ratings across ratees for each category in each cell, and (2) the corresponding one-rater validity coefficients employing Guilford's (1954, p. 407) equation for adjusting validity in terms of changes in test length. (In this case, the adjustment was from 24 raters[2] to one rater.) The resulting four sets of correlation

---

[2]The actual step-down factor varied from 24 in some cases due to missing data. The exact step-down factor employed was Ebel's (1951, p. 413) k_o.

coefficients were then transformed for use in analyses of variance using the Pearson r-to-Z transformation (Guilford & Fruchter, 1978, p. 522).

The general hypotheses examined in this study were:

H_8 Training and Participation have significant main and interactive effects on r-to-Z transformed intraclass reliability coefficients. (Training and Participation are expected to result in significantly greater transformed reliability coefficients.)

H_9 Training and Participation have significant main and interactive effects on r-to-Z transformed one-rater reliability coefficients. (Training and Participation are expected to result in significantly greater transformed reliability coefficients.)

H_{10} Training and Participation have significant main and interactive effects on r-to-Z transformed intraclass validity coefficients. (Training and Participation are expected to result in significantly greater transformed validity coefficients.)

H_{11} Training and Participation have significant main and interactive effects on r-to-Z transformed one-rater validity coefficients. (Training and Participation are expected to result in significantly greater transformed validity coefficients.)

Each of the above general hypotheses was tested using an individual ANOVA design. Again, the nature of the interaction was not specified. However, it was expected that those subjects who had both participated in developing the rating instrument and been trained in its use would produce the most reliable and valid ratings.

Phase Two: Effects of Treatments on Mediators

The hypothetical model diagrammed in Figure 1 specifies that both rater participation in scale construction and rater training should have measurable effects on raters' attitudes toward performance appraisal rating, and that rater training should also influence raters' knowledge

of the performance appraisal rating process. The two studies included in this phase of the investigation were designed to examine these aspects of the model. Possible interactions between the treatments, as well as their main effects, were evaluated in these two studies.

Study Six. This study was intended to measure the effects of Training, Participation, and the Training x Participation interaction on attitudes toward performance appraisal rating as operationalized in Chapter III. The general hypothesis examined here was:

> H_{12} Training and Participation have significant main and interactive effects on Attitude. (Training and Participation are expected to result in significantly higher levels of Attitude.)

This hypothesis was tested using an ANOCOV design with Participation and Training serving as the independent variables, post-treatment Attitude scores as the dependent variable, and pre-treatment Attitude scores as the covariate variable.

Study Seven. The purpose of this study was to examine the main and interactive effects of Training and Participation on raters' knowledge of the performance appraisal rating process, as operationally defined in Chapter III. The general hypothesis examined in this study was:

> H_{13} Training has a significant main effect on Knowledge, but Participation has no such effect, and there is no significant interaction effect. (Training is expected to result in a significantly higher level of Knowledge.)

This hypothesis was tested using an ANOCOV design with Participation and Training serving as the independent variables, post-treatment Knowledge scores as the dependent variable, and pre-treatment Knowledge scores as the covariate variable.

Phase Three: Effects of Treatments on Outcomes with Mediators Covaried

The model suggests that Participation and Training are effective in changing characteristics of ratings solely because of the effects of the two treatments on the two mediating variables. If this aspect of the hypothetical model is correct, then by holding any changes in the mediating variables due to the treatments constant through analysis of covariance, any statistically significant effects of the treatments on outcomes discovered in Phase One should be made to disappear. The three studies included in this section, corresponding to the first three studies described in Phase One, were intended to investigate this implication of the hypothetical model.

Study Eight. This study was intended to examine the main and interactive effects of Training and Participation on the overall elevation of the ratings when the effects of the treatments on the hypothesized mediators were accounted for. (This analysis corresponds to the "between subjects" portion of Study One.) The general hypothesis under consideration in this study was:

> H_{14} There are no significant main or interactive effects of Training and Participation on overall elevation of the ratings when changes in Attitude and Knowledge scores due to the effects of the treatments are accounted for.

This rather complex general hypothesis was tested using three ANOCOV designs. The first ANOCOV employed two treatments as independent variables, ratings on each ratee for each category as the dependent variable, and change scores in Attitude as the covariate variable. The second ANOCOV used scores in Knowledge as the covariate variable. Change scores in both Attitude and Knowledge were employed as covariate

variables in the third ANOCOV. By sequencing the analyses in this manner, information was obtained to help support or refute the hypothetical model diagrammed in Figure 1.

Study Nine.[3] The purpose of this study was to have been to examine the effects of Training, Participation, and the Training x Participation interaction on elevation of ratings per category when the effects of the treatments on the hypothesized mediators were accounted for. (This analysis corresponds to Study Two.) The general hypothesis to have been examined here was:

H_{15} For each category, there are no significant main or interactive effects of Training and Participation on elevation of the ratings when changes in Attitude and Knowledge scores due to the effects of the treatments are accounted for.

Study Nine was to have employed a MANOCOV design similar to that used in Study Eight.

Study Ten. This study was intended to examine the main and interactive effects of Training and Participation on dispersion of ratings per category when the effects of the treatments on the hypothesized mediators were accounted for. (This analysis corresponds to Study Three.) The following general hypothesis was under examination here:

H_{16} For each category, there are no significant main or interactive effects of Training and Participation on dispersion of the ratings when changes in Attitude and Knowledge scores due to the effects of the treatments are accounted for.

---

[3] Study Nine was not actually performed, since the hypotheses tested in Study Two were not supported.

Again, a sequence of three MANOCOVs, using Attitude change scores, Knowledge change scores, and Attitude and Knowledge change scores respectively as covariate variables, were employed to test this hypothesis. The MANOCOVs were identical to the MANOVA used in Study Three except for the addition of the covariate variables. When warranted by the findings of the MANOCOVs, individual ANOCOVs for each category were interpreted.

CHAPTER III

METHOD OF INVESTIGATION

Subjects

The Experiment Proper

Undergraduate students from over 20 classes in introductory, developmental, and social psychology, statistics, death and dying, and the psychology of women were solicited to serve as subjects in this experiment.[1] The experimenter informed potential subjects of the general nature of the experiment and offered them five percentage points of extra credit in their psychology classes plus an approximate one-in-twenty-five chance at winning a $50 bill (one lottery per experimental cell) as compensation for their participation. Potential subjects were asked to sign a sheet of paper and provide their telephone numbers if they were interested in participating in the study. One-hundred and fifty students expressed interest. These potential subjects were randomly assigned to the four cells of the experimental design: (a) Both Participation and Training, (b) Participation Only, (c) Training Only, and (d) Neither Participation nor Training. They were then telephoned and told where and when to report. During this telephoning process several subjects withdrew their statements of interest, and others expressed a

---

[1]The author is indebted to the Auburn University Psychology Department's Faculty-Student Ethics Committee for carefully reviewing the experiment in its proposed form and providing written approval for it to be carried out with Auburn University students as subjects.

need to be changed to another group due to meeting-time conflicts with classes, laboratories, and personal commitments. One-hundred and forty-five subjects, assigned to cells as noted in Table 2, survived the telephoning phase of subject solicitation.

One-hundred subjects, 25 per cell, attended the first experimental session. Through the use of telephone reminders and make-up sessions, 97 subjects were retained through the entire five-week duration of the experiment. The three subjects who were lost (one per cell) either resigned from school or withdrew from their psychology classes and were no longer in need of extra credit. Data for one randomly-chosen subject in Group D, Neither Participation nor Training, were dropped from the analyses. Thus, the statistical analyses were performed using a subject sample size of 96, with 24 subjects per cell.

Subjects were clearly informed at the beginning of the first experimental session of the nature of the study, their compensation, and their rights and obligations as experimental subjects. All subjects read and signed the Program Evaluation Study Informed Consent Form (Appendix A-1) at the first session, and were thoroughly debriefed at the conclusion of the experiment.

Related Questionnaire Development Studies

Six questionnaire development studies, described later in this chapter, were carried out in order to construct and evaluate the measurement instruments and rating scales employed in the experiment. As tabulated in Table 3, a total of 350 subjects participated in these six studies. These subjects were students enrolled in introductory, social

Table 2. Subject Participation in the Experiment Proper

Phase	Group			
	A. Both Participation and Training	B. Participation Only	C. Training Only	D. Neither Participation nor Training
Originally assigned	38	37	37	38
After telephoning	39	33	35	35
Session One				
Regular	24	24	23	25
Make-up	1	1	2	0
Session Two				
Regular	21	22	21	22
Make-up	3	2	3	3
Session Three				
Regular	20	24	20	22
Make-up	4	0	4	3
Session Four				
Regular	22	20	19	23
Make-up	2	4	5	2
Session Five				
Regular	23	22	19	23
Make-up	1	2	5	2
Included in Analysis	24	24	24	24[a]

[a]Data for one randomly-selected subject in Group D were dropped from the analysis in order to allow equal cell sizes.

Table 3. Subject Participation in the Questionnaire Development Studies

Study Number	Purpose	Number of Subjects Participating[a]	Duration of Study
1	Provide ideas for attitude questionnaires via essays about teacher performance evaluation.	36	1 hour
2	Evaluate 175 attitude statements; provide test-retest reliability data for the 150-item Knowledge Scale.	115	4 or 6 hours[b]
3	Evaluate 75 attitude statements.	50	1 hour
4	Evaluate 75 additional attitude statements.	50	1 hour
5	Provide test-retest reliabilities for the 150-item Knowledge Scale, 30-item Attitude Scales A and B, and 1-item Criterion Scale; provide parallel-form reliability and concurrent validity estimates for Attitude Scales; provide convergent and discriminant validity estimates for the set of scales.	47	1 or 3 hours[b]
6	Evaluate 579 critical incidents of professor behavior. (Maximum of 150 incidents evaluated by each subject.)	52	1 hour

[a]A total of 350 subjects participated in these six studies.

[b]Subjects who completed both sets of forms in the test-retest portion of the study received the larger number of hours credit. Those who completed only "test" or "retest" forms received the smaller number of hours credit.

and industrial psychology and human sexuality classes.[2] All subjects in the questionnaire development studies were informed of the nature of the studies and were required to read and sign an appropriate version of the Questionnaire Development Study Informed Consent Form (Appendix A-2). Subjects were compensated with the number of extra credit points in the psychology classes appropriate for the amount of time they spent in the various questionnaire development tasks (one-half percentage point per hour of participation). All subjects were given an opportunity to attend debriefing sessions held at the end of the experiment.

Instruments and Materials

Attitude Scales

Subjects' attitudes toward the performance appraisal rating process were measured using two 30-item Thurstone-type (equal-appearing interval) scales (Edwards, 1957, pp. 83-119; Thurstone & Chave, 1929): Forms A and B (see Appendices B-1 and B-2). Form A was administered as a pretest, Form B as a posttest. The score for each form is the mean scale value of the items with which the subject agrees (circles "A"). The forms were developed by the experimenter for use in this project. Reliability estimates for the scales, obtained in the manner described later in this chapter, ranged from .58 to .81. Estimates of concurrent validity with a criterion measure (see Appendix D) ranged from .65 to .81.

---

[2]The Auburn University Psychology Department's Faculty-Student Ethics Committee also provided written approval for these studies to be carried out using Auburn University students.

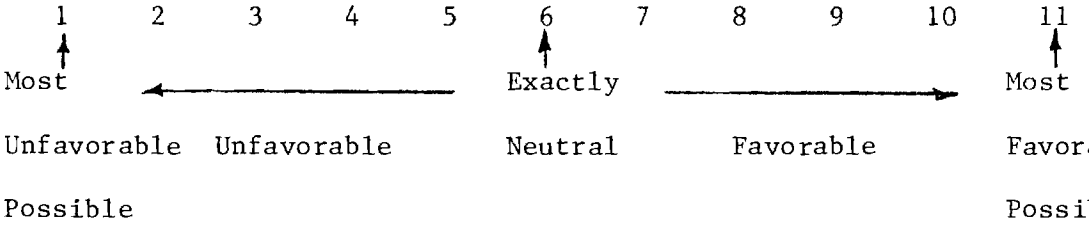
The scale development process, following the general guidelines specified by Edwards (1957, pp. 83-119), Selltiz et al. (1976, pp. 414-417), and Thurstone and Chave (1929), progressed as follows:

1. A group of 36 students was asked to write short essays about the process of evaluating the classroom teaching performance of college professors. (This is the first Questionnaire Development Study referred to above.) The assignment was deliberately loosely defined: In their essays, the students mentioned such things as their own feelings about the rating process in general, the shortcomings and strong points of the evaluation forms they had used, their own suggestions for teacher evaluation, and problems they had encountered with various professors, courses, and/or rating forms. The subjects were asked to formulate at least three favorable and three unfavorable statements about the process of evaluating college classroom teaching performance.

2. From these essays and the writings of several experts in the field of personnel evaluation (Baridon & Loomis, 1931, pp. 161-201; Famularo, 1972, pp. 40:1 - 45:17; Halsey, 1953, pp. 133-149; Kellogg, 1965; Kornhauser, 1926; Luck, 1955; Rowland, 1970), a number of statements expressing favorable and unfavorable attitudes toward the performance appraisal process were drawn. After editing and revision, a set of 175 statements was put into questionnaire form to be evaluated by subjects similar to those who would eventually be asked to respond to them in the experiment.

3. One-hundred and fifteen students evaluated each of these 175 statements in terms of the degree of favorability/unfavorability they

expressed toward the performance appraisal rating process. (This task was included in Questionnaire Development Study Two mentioned above.) The evaluators used an 11-point scale anchored as follows:



The evaluators were given detailed instructions in how to carry out the item ratings, and were provided with two examples. The mean and standard deviation of the evaluative ratings for each instrument were calculated and used as guideposts to determine whether or not to consider each item for inclusion in a form of the actual attitude questionnaire. Many of the items were abandoned due to the high standard deviation of their ratings (indicating that the subjects could not agree on the favorability of the item); others were rewritten or revised. The majority of the items were found to express somewhat neutral opinions (mean ratings of four to eight). Fifty of the more discriminating items were combined with 25 new items written by the experimenter to form a new set of statements for students to evaluate.

4. Fifty students evaluated the new list of 75 attitude statements using the 11-point scale presented above. (This is Questionnaire Development Study Three.) As before, detailed instructions and examples were provided. The resulting item ratings were tabulated, and Thurstone's S and Q values (see Edwards, 1957, pp. 86-89) were calculated for each

item.[3] An examination of the S and Q values of the 75 items indicated that a number of the items might be suitable for inclusion in an attitude scale. However, there was a paucity of items at the far extremes of the favorableness-unfavorableness continuum, as well as near the "exactly neutral" point. For this reason, an additional 75 statements were prepared by the experimenter and formed into another questionnaire for students to evaluate.

5. Again, 50 students evaluated these 75 items using the 11-point scale after receiving detailed instructions including examples. (This is Questionnaire Development Study Four.) Thurstone's S and Q values were calculated for these items as well.

6. The 150 items evaluated in steps (4) and (5) were ordered on a favorability-unfavorability continuum by means of their S values. The six best items (in terms of Q value; the smaller the better) within each S value level were retained for inclusion in the final versions of the attitude questionnaire. When six items were not available for an S value level, all items with Q values less than 2.0 were retained. Sixty items were chosen to form the final versions of the scale. The items were again listed in order by S value and were assigned to scale forms using the ABBAABB...A procedure recommended by Thurstone and Chave (1929, p. 65). The resulting prototype forms were examined carefully; one of any pair of items within the same form which appeared

---

[3] "S," the median rating, defines the scale value of the item; "Q," the semi-interquartile range, represents the extent of agreement among raters concerning the scale value.

redundant was switched with an item from the other form with similar S and Q values. Items were ordered randomly within the two final versions of the questionnaire. The final versions of the two questionnaires, along with their item statistics, are found in Appendix B.

7. The two forms of the Attitude Scale were each pilot-administered twice prior to their use in the experiment. The test-retest reliability estimates of the two forms--.78 for Form A, .79 for Form B--along with other data described below, are displayed in Table 5. Means and standard deviations, respectively, for the Attitude Scale Form A were 7.43 and 0.91 (pretest), 7.48 and 0.89 (posttest); for Form B they were 7.05 and 0.87 (pretest), 7.11 and 0.80 (posttest). The two forms of the Attitude Scale were not strictly parallel in terms of means, despite attempts to make them so during the scale development process. This fact is taken into consideration in later chapters when the results of studies employing these instruments are presented and discussed. The base attitude level of the subject sample appeared to be slightly favorable toward the use of teacher performance appraisal instruments.

Knowledge Scale

Subjects' knowledge of the performance appraisal rating process was measured with a 50-item true-false test designed by the experimenter to assess knowledge of the content domain covered in the rater training program[4] incorporated in this experiment. These 50 relevant items

[4]The rater training program is described in detail later in this chapter. A copy of the outline of the program is attached as Appendix H.

are randomly interspersed among 100 additional items measuring knowledge of two irrelevant (for the purposes of this experiment) content areas: (a) the history and policies of Auburn University, as presented in the 1977-1978 Auburn University Bulletin; and (b) practical learning methods, as presented in a handout prepared by the experimenter for use in an introductory psychology class. The 100 irrelevant items were included as fillers in order to (a) partially mask the true purpose of the test; (b) guard against memorization of the items; (c) alleviate, as much as possible, sensitization to the content to be covered in the rater training program; and (d) provide a control measure to test the Hawthorne effect. The Knowledge Scale was employed as both a pretest and a posttest, thus these special precautions were necessary. A subject's score is simply the number of items he answers correctly. The test is displayed in Appendix C.

The first step in the development of the Knowledge Scale was the preparation of a detailed outline of the training program (see Appendix H). Items were then written such that each major point of the outline was represented by at least one item. More important points, such as III-A-4, which deals with common biasing factors affecting ratings, were given more emphasis through representation by several items. (In the specific case of point III-A-4, nine items were written.) The 50 items are thus designed to represent fairly well the major content areas of the training program.

In order to evaluate the content validity of the test, five judges were given copies of the test and the training program and were asked to

identify any test items which were not representative of some aspect of the training program. All five judges agreed that all 50 items were relevant. In a more strenuous test of content validity, the five judges were asked to identify the exact point on the training program represented by each item (by placing test item numbers in blanks drawn next to the training program outline points). The judges' agreement with the experimenter's key of item placements was as follows: Rater A, a sophomore undergraduate majoring in hospital administration, agreed 88% with the experimenter; Rater B, a second-year graduate student majoring in industrial psychology, agreed 96% with the experimenter; Raters C, a first-year graduate student majoring in clinical psychology, D, a second-year graduate student majoring in industrial psychology, and E, a Research Associate investigating energy consumption issues, agreed 100% with the experimenter.

Several studies were performed to assess the reliability of the Knowledge Scale. Reliability coefficients ranged from .64 to .75, as shown in Table 4. Apparently the base knowledge level of the subject sample was quite high even without specific training, since the mean scores presented in Table 4 are much higher than the score of 25 expected by chance.

A Study of the Reliability and Validity of the Attitude and Knowledge Scales

Questionnaire Development Study Four represents the major effort to assess the convergent and discriminant validities of the Knowledge and Attitude Scales. The Knowledge Scale and Attitude Scale Forms A

Table 4. Means, Standard Deviations, and Reliability Estimates for the Knowledge Scale

Sample	n	Mean	Std. Dev.	Reliability Estimate	r
Questionnaire Development Study Two, Pretest	90	39.14	4.87	Kuder-Richardson 20	.71
Questionnaire Development Study Two, Posttest	62	40.62	4.12	Kuder-Richardson 20	.64
Questionnaire Development Study Four, Pretest	43	39.16	4.87	Kuder-Richardson 20	.71
Questionnaire Development Study Four, Posttest	40	40.25	4.34	Kuder-Richardson 20	.67
Questionnaire Development Study Two	37	38.49[a] 40.49[b]	5.09[a] 4.49[b]	One-week Test-Retest	.73
Questionnaire Development Study Four	36	39.36[a] 40.17[b]	4.94[a] 4.52[b]	One-week Test-Retest	.75

[a]Pretest statistics.

[b]Posttest statistics.

and B, along with the Criterion Scale, an 11-point single-item rating scale of overall attitude toward the use of rating forms for evaluating college teaching performance (see Appendix D), were administered twice, separated by a one-week interval, to a sample of 47 undergraduate students. The correlation coefficients resulting from this study are presented in Table 5. Correlations for the pre- and posttest administrations of the same scale represent test-retest reliability (stability) estimates. Correlations for Attitude Scale Forms A and B within the same administration represent alternate-forms reliability (equivalence) estimates. Correlations between Attitude Scale Forms A and B and the Criterion Scale within the same administration represent concurrent validity estimates.

It is important to note that all of the measures dealing with attitude (Attitude A Pretest and Posttest, Attitude B Pretest and Posttest, and Criterion Pretest and Posttest) are significantly intercorrelated (see Guilford, 1965, pp. 580-581), and the two measures of knowledge (Knowledge Pretest and Posttest) are significantly intercorrelated, but there are no significant correlations between measures of attitude and knowledge. Claims for the construct validity of the two sets of measures on the basis of the convergent-discriminant validity criteria (Campbell & Fiske, 1959) appear justified.

Behaviorally Anchored Rating Scales (BARS)

The Behaviorally Anchored Rating Scales (BARS) for evaluating the effectiveness of instructors in terms of five categories of college classroom teaching performance, as employed in this investigation, are

Table 5. Intercorrelations and Reliability Coefficients of the Attitude, Knowledge, and Criterion Scales

Variable	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Knowledge Pretest	.24 (43)	.19 (43)	.28 (43)	.75** (36)	.23 (36)	.18 (36)	.28 (36)
(2) Attitude A Pretest		.81** (43)	.72** (43)	.42 (36)	.78** (36)	.73** (36)	.76** (36)
(3) Attitude B Pretest			.74** (43)	.36 (36)	.71** (36)	.79** (36)	.80** (36)
(4) Criterion Pretest				.35 (36)	.81** (36)	.65* (36)	.92** (36)
(5) Knowledge Posttest					.31 (40)	.31 (40)	.37 (40)
(6) Attitude A Posttest						.58* (40)	.81** (40)
(7) Attitude B Posttest							.67** (40)
(8) Criterion Posttest							

Note. The number in parentheses below each correlation coefficient is the number of pairs (n) on which the coefficient was calculated. All significance levels are taken from Guilford's (1965, pp. 580-581) Table D, which takes into account degrees of freedom and number of variables.

*$p < .05$

**$p < .01$

displayed in Appendix F. For each category, respondents simply place an "X" at the point along the 11-inch vertical graphic scale which they feel best represents the level of performance exhibited by the instructor they are evaluating. The ratings are scored to one decimal place by means of an engineer's scale, which divides an inch into ten equal units.

The BARS were developed during the course of this investigation. The actual development process, following generally the guidelines specified by Bernardin, LaShells, Smith, and Alvares (1976), Campbell et al. (1973), and Smith and Kendall (1963), was incorporated as one treatment in the experiment. The scale construction process progressed as follows:

1. Meeting together as one group, the 48 subjects assigned to cells A (Both Participation and Training) and B (Participation Only) discussed, under the leadership of the experimenter, the nature of the college classroom teacher's duties and behavior, and generated in a brainstorming session a list of approximately 20 identifiable categories of teacher behavior which might be evaluated. During the classroom period which immediately followed, the list was refined and reduced to the five categories of behavior which the group felt were most relevant to effective college classroom teaching and were most easily identifiable and measurable. These categories, and their definitions, are presented in Table 6. The categories are similar to those identified in previous research projects involving college classroom teaching effectiveness evaluation (e.g., Friedman & Cornelius, 1976; Harari & Zedeck, 1973; Keaveny & McGann, 1975; Ronan, 1971, 1972; Zedeck, Jacobs, & Kafry, 1976).

Table 6. Definitions of the Five Categories of College Classroom Teaching Behavior

A. Relationships with Students. This category refers to the way the professor treats his/her students both in and out of class. It includes such things as talking with students before, during, and after class, interacting with and counseling students in the office and elsewhere regarding course-related and personal problems, knowing students' names, and treating students with respect in class.

B. Ability to Present the Material. This category refers to the way the professor organizes the material and presents it to the class. It includes such things as coming to class well-prepared and on time, organizing the material in a logical manner, speaking and writing clearly, and using examples, audio-visual aids, and other devices to get the material across to the students.

C. Interest in Course and Material. This category refers to the professor's knowledge of and interest in the material he/she is trying to teach. It includes such things as being able to answer questions and elaborate on the material, showing enthusiasm for the course, and reading and researching to keep current and learn more about the subject matter.

D. Reasonableness of the Workload. This category refers to the amount of work (reading, homework problems, class and lab work, papers, tests, etc.) assigned by the professor. It includes such things as clearly specifying assignments and due dates, scheduling the work evenly throughout the quarter, and keeping the workload appropriate to the credit-hour value of the course.

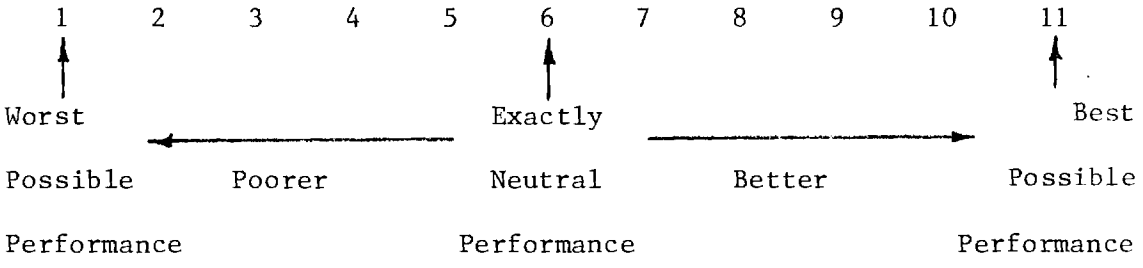
E. Fairness of Testing and Grading. This category refers to the fairness of the professor's testing and grading policies. It includes such things as stating how grades are to be determined, testing over appropriate material, and grading without bias.

At the conclusion of this first session, subjects were told that they would be asked to provide critical incidents of effective, mediocre, and ineffective behavior for each of the five dimensions they had identified.

2. One week following the first session described above, the 48 group A and B subjects again met together as a group. During this second session they each provided three critical incidents of teacher behavior (one effective, one mediocre, one ineffective), which they had themselves observed, for each category identified in the first session. These incidents were written by the subjects on a form provided by the experimenter (see Appendix E-1).

3. The experimenter read the 720 incident descriptions written during session two and culled duplicates, vague statements, and descriptions which could not really be classified as incidents. The remaining incident descriptions were edited by the experimenter into one-sentence statements which summarized the descriptions provided by the subjects. The actual language used by the subjects was preserved as closely as possible. Five hundred and seventy-nine incident statements survived this editing process. These incident descriptions were randomly ordered and distributed among four forms of an item evaluation questionnaire. The instructions for the four forms of the questionnaire were identical: Respondents were to read each incident statement and determine, using the set of category definitions provided in the questionnaire, which category each statement represented. Respondents were also to rate each incident statement on an 11-point scale of performance effectiveness

anchored as follows:



Respondents were provided with detailed directions and two worked examples on the instruction sheet.

4. During session three, held one week following the second session, the 48 subjects in groups A and B evaluated the 579 incidents using the four forms constructed in step (3). Twelve copies of each form were distributed randomly among the 48 experimental subjects, thus no subject was required to evaluate more than 150 incidents. So that 25 ratings would be available for each incident, an additional subject sample of 52 undergraduate students was also asked to fill out the evaluation forms. (This was Questionnaire Development Study Six mentioned earlier in this chapter.) As with the experimental subjects, 13 copies of each form were distributed randomly among these 52 evaluators such that no subject was required to evaluate more than 150 incidents.

5. The 25 category placements and effectiveness ratings collected for each of the 579 incident statements were tabulated by the experimenter with the assistance of several colleagues. Three statistics were calculated for each incident: (1) percentage of agreement for placement of the item in the modal category; (2) median effectiveness rating of the incident (S value); (3) semi-interquartile range of the

effectiveness ratings for the incident (Q value). Items were selected for inclusion in a potential scale anchor pool on the basis of two criteria: (1) percentage of agreement for category placement not less than 60% (see Bernardin, LaShells, Smith, & Alvares, 1976); (2) Q value not greater than 3.0. Four hundred and forty-three incidents were included in the pool. Most of the 443 incidents bettered the two criteria by a substantial margin. The incidents were then sorted into categories and S-value levels within category. Despite the attempt to collect incidents reflecting mediocre performance, as well as those representing effective and ineffective behavior, the number of incidents judged to represent the neutral areas of several categories were rather sparse.[5] However, there were enough items to adequately anchor a set of BARS.

6. Appropriate anchors for each category were selected from the item pool, rewritten into Smith and Kendall's (1963) "expectation" format, and placed along 11-inch-long vertical graphic rating scales by means of an engineer's scale. The resulting BARS for each of the five categories, along with their item statistics, are presented in Appendix F. The actual order of categories was randomized for each individual BARS booklet used in this investigation.

Simulated Professors

A potential problem in the interpretation of the results of many studies of performance appraisal ratings is that the stimuli being rated are not standardized. For example, subjects in several work groups or

---

[5]This is a common problem with the BARS development process. See, for example, Harari and Zedeck (1973), Landy and Guion (1970), and Zedeck et al. (1974).

classes may be asked to evaluate their most recent supervisors or instructors, yet the particular individuals being rated may vary greatly across many variables (see Category V in Table 1). Some experimenters choose to assume away these differences, others attempt to control them through random assignment of subjects to groups or through matching techniques. There is no assurance, however, that such procedures do in fact adequately control ratee differences. When the actual characteristics of the ratees are uncontrolled, it is difficult to determine whether, for example, mean differences in ratings are due to the leniency error or to real differences among ratees. Even when subjects are asked to evaluate the same individuals, their opportunity to observe those individuals may vary. Recognizing these potentially uncontrolled sources of variance, several investigators (e.g., Borman, 1978; Lifson, 1953; Wexley et al., 1972, 1973) have attempted to provide all raters with identical stimuli during their experimental studies of various rating phenomena.[6]

An attempt was made in this investigation to control sources of variance due to ratees, and to opportunities to observe ratee behavior, through the use of a set of standardized, simulated professors as stimuli for appraisal. The stimuli to be rated actually consisted of five descriptions of imaginary college professors, labeled L, M, N, O, and P, created by the experimenter. Each description was followed by a simulated

---

[6]It must be recognized that this approach may sacrifice generalizability in order to obtain experimental control.

behavioral incident diary (Bernardin & Walter, 1977; Bernardin & Boetcher, Note 9), constructed by the experimenter, of 20 statements taken from the pool of scaled incident descriptions created in step (5) of the BARS construction process described above. Care was taken to ensure no duplications of incidents among simulated diaries or between simulated diaries and BARS. The experimenter was also careful to ensure reasonable consistency among behavioral incidents included within each simulated diary. Statements were chosen such that each category of behavior was represented by four behavioral incidents; however, the order of incident description statements was randomized within each diary. Incidents were selected and assigned to simulated professors on the basis of $\underline{S}$ value so that each professor's "true score" on each category was known.[7]

Additionally, an attempt was made to assign incidents in such a manner that the statistical analyses described in the next chapter could most easily be interpreted. Ease of interpretation would be facilitated by assigning to professors incidents with specified scale values such that the row sums and column sums of the professor x category matrix are constant. Due to limitations of the incident pool, this goal was not completely possible. However, as can be seen from Table 7, the row and column sums, means, and variances were nearly enough equal that interpretation of the results of the various analyses involving

---

[7]Note that the use of these standardized diaries eliminated such sources of variance in ratings as the amount of information to which the rater is exposed and the rater's freedom to select which behaviors to observe.

Table 7. Scale Values in the Simulated Professor x Category Matrix

Category	Simulated Professor L	M	N	O	P	Row Sum	Row Mean	Row Variance
A	4.0	10.0	8.0	2.0	6.0	30.0	6.0	8.0
B	6.0	2.0	10.0	4.0	8.0	30.0	6.0	8.0
C	8.0	4.0	6.0	10.0	2.0	30.0	6.0	8.0
D	9.6	6.0	2.0	8.0	4.0	29.6	5.9	7.4
E	2.0	8.0	4.0	6.1	10.0	30.1	6.0	8.0
Column Sum	29.6	30.0	30.0	30.1	30.0			
Column Mean	5.9	6.0	6.0	6.0	6.0			
Column Variance	7.4	8.0	8.0	8.0	8.0			

the ratings were greatly facilitated. The five simulated professors and their respective behavioral incident diaries are attached as Appendix G. Appendix G also contains the item statistics for the incident descriptions contained in each simulated diary.

The actual order of presentation of the simulated professors was randomized for each individual subject. All subjects evaluated exactly the same stimuli, yet other potentially biasing factors may be included in the descriptions of the simulated professors. Such other biases may be manifested as differences in mean ratings across professors and categories.

Training Program

A detailed outline of the training program is provided as Appendix H. The outline follows the sequence specified by Bittner (1948, pp. 425-426; see Chapter I) and includes the three major components recommended by Brown (1968): (a) practice with the specific scales used in the rating program (Bernardin & Walter, 1977; Wakely, Note 11); (b) discussion of rating errors (Latham et al., 1975; Levine & Butler, 1952); and (c) special emphasis on the importance of trait differentiation (Latham et al., 1975; Taylor & Hastman, 1956). The content of the training program, developed by the experimenter, draws heavily from material supplied in the writings of Baylie et al. (1974), Bittner (1948), Guilford (1954, pp. 263-301), and Smith (1976). The training program presented in the experiment consisted of lecture, discussion, supervised practice, and two readings: a copy of the training program outline and a set of the BARS constructed in the first three experimental sessions.

Treatment Conditions and Procedure

As noted in the "Subjects" section of this chapter, subjects were assigned to four treatment conditions: (a) Both Participation and Training, (b) Participation Only, (c) Training Only, and (d) Neither Participation nor Training (Control). This experiment was carried out over a period of five weeks, with experimental sessions conducted on Monday afternoons, 4:00 to 6:00 p.m., and control sessions conducted on Tuesday afternoons, 4:00 to 6:00 p.m. Make-up sessions for those few subjects who were forced to miss their regularly scheduled sessions were held at times mutually convenient to the subjects and the experimenter. The procedural plan of the experiment is diagrammed in Figure 2. The experimental and control sessions are described below. Note that two groups were scheduled to attend each session; the two groups for each session met jointly to ensure identical treatment.

Experimental Sessions

Session E1. Subjects were informed of the general nature of the experiment and were required to read and sign an informed consent form (see Appendix A-1). They were also specifically instructed not to discuss the experimental procedures with anyone until the end of the fifth session. They were then administered the Knowledge Scale and Form A of the Attitude Scale. When everyone had completed the scales, the procedure described as step (1) of the BARS development process commenced. Subjects discussed the evaluation of college teaching performance, generated approximately 20 categories of teacher behavior, and refined the list to five major categories. The session concluded with the

	Week 1	Week 2	Week 3	Week 4	Week 5
Group A: Both Participation and Training	Session E1	Session E2	Session E3	Session E4	Session E5
Group B: Participation Only	Session E1	Session E2	Session E3	Session C4	Session C5
Group C: Training Only	Session C1	Session C2	Session C3	Session E4	Session E5
Group D: Neither Participation nor Training	Session C1	Session C2	Session C3	Session C4	Session C5

Figure 2. Procedural Plan of the Experiment

experimenter's request that subjects try to think of three critical incidents (one effective, one mediocre, one ineffective) for each category of behavior.

Session E2. This was the second BARS development session as described in step (2) of the rating scale development procedure. Using the critical incident reporting form displayed as Appendix E-1, subjects (anonymously) supplied an effective, a mediocre, and an ineffective critical incident for each teacher behavior category.

Session E3. This was session three, described in step (4), of the BARS development procedure. Using the four forms of the item evaluation instrument, subjects categorized and ranked in terms of effectiveness a maximum of 150[8] edited incident statements.

Session E4. Subjects in this session were exposed to the rater training program as described above. They were provided copies of the training program outline (Appendix H) and BARS (Appendix F), to which they referred during the lecture and discussion portions of the training program. The discussion of various rating errors was accompanied by illustrations on the blackboard. Subjects were instructed to use the BARS to evaluate the performance of at least one of their previous instructors (to be kept anonymous) during the last phase of the training program. They were also invited to ask questions about anything they did not understand, and to take the training program outline and BARS home with them for future reference. Subjects were strongly encouraged

---

[8]One form of the item evaluation instrument contained only 129 incident description statements.

to practice evaluating additional instructors during the interval between sessions.

Session E5. After a brief presentation of instructions concerning the use of BARS, subjects evaluated the five simulated professors with the BARS. Note that the order of presentation of the simulated professors was randomized for each subject and that the order of categories was randomized within each BARS booklet. When all subjects were finished with their rating task, they were administered the Knowledge Scale and Form B of the Attitude Scale. When all scales were completed, the lottery for the $50 bills was held, and subjects were completely debriefed, released from their commitment not to discuss the experiment, and dismissed.

Control Sessions

Session C1. Again, subjects were informed of the general nature of the experiment and were required to read and sign an informed consent form (see Appendix A-1). As in Session E-1, subjects were specifically instructed not to discuss the experimental procedures with anyone until the end of the fifth session. They were then administered the Knowledge Scale and Form A of the Attitude Scale. When everyone had completed the scales, subjects were informed that the next several sessions would involve experiences of adolescents. Following a short discussion of the period of adolescence and the types of activities important during this period, subjects were told that during the next two sessions they would be asked to (anonymously) supply critical incidents exemplifying times that they felt good and bad during their own adolescence.

Session C2. Using the critical incident form displayed as Appendix E-2, subjects were asked to (anonymously) supply descriptions of at least 15 incidents which made them feel "really good" during adolescence. They were informed that other students might see their anonymous incident descriptions in later session, and were told to specify any incident descriptions that they would prefer not be made public.

Session C3. Using the critical incident form displayed as Appendix E-3, subjects were asked to (anonymously) provide descriptions of at least 15 incidents which made them feel "really bad" during adolescence. Again, they were informed that other students might see their incident descriptions, and were instructed to specify any descriptions they wished kept confidential.

Session C4. Subjects were assigned to small groups (approximately six subjects per group). Each group was provided with approximately 150 of the critical incident descriptions[9] generated in Sessions C2 and C3--some relating to good feelings, some to bad. Each group was asked to review the incident descriptions and to generate lists of categories of incidents which led to good and bad feelings during adolescence. When all groups had completed these lists, master lists were constructed on the blackboard from the oral inputs of all the small groups.

Session C5. Session C5 was identical to Session E5.

---

[9]Incident descriptions requested to be kept confidential were not employed in this session.

CHAPTER IV

RESULTS

The data analysis was carried out in three phases. Phase One, containing studies One, Two, Three, Four, and Five, was intended to examine the effects of Participation and Training on several aspects of the ratings resulting from the procedure discussed in the previous chapter. Phase Two, consisting of Studies Six and Seven, was designed to determine whether Participation and Training affect Attitude and Knowledge. Studies Eight, Nine, and Ten, forming Phase Three of the analysis, were intended to explore the relationship among the treatments, mediators, and outcomes. Specifically, they were designed to determine whether any statistically significant findings in several of the Phase One analyses remained when the effects of Participation and Training on Attitude and Knowledge were held constant through use of analysis of covariance.[1]

Study One

Purpose

The purpose of this study was to examine the effects of Participation and Training on the ratings obtained during the fifth session of

---

[1] All major statistical analyses were performed on Auburn University's primary computer, an IBM System 370 model 158. The various programs included in the Statistical Analysis System (SAS; Barr & Goodnight, 1972; Barr, Goodnight, Sall, & Helwig, 1976; Helwig, 1977) were employed whenever possible. Other analyses were performed by the experimenter using a Texas Instruments SR-51A calculator.

the experiment. The four general hypotheses examined in this study are specified in Chapter II. The analysis employed was a split-plot factorial ANOVA (see Kirk, 1968, pp. 311-312) with Participation and Training (two levels each) serving as between-subjects independent variables and Categories and Professors (five levels each) as within-subjects independent variables. The ω^2 statistic (see Kirk, 1968, pp. 198-199) was calculated to determine the practical significance of any sources of variance found statistically significant in the ANOVA.

Findings

The ANOVA table is presented in Table 8. Statistically significant (at $\alpha = .05$) sources of variance are Training, Categories, Professors, Participation x Professors, Training x Professors, Categories x Professors, and Participation x Categories x Professors. In order to interpret the various interactions with Participation and Training, four additional ANOVAs were performed. Tables 9 and 10 summarize the results of ANOVAs carried out on two subgroups of the sample--Participant subjects (those in experimental cells A and B) and Non-participant subjects (those in experimental cells C and D) respectively. Individual ANOVAs were also carried out for Trained subjects (cells A and C) and Untrained subjects (cells B and D); these are presented in Tables 11 and 12 respectively.

The results indicate that Training significantly decreased the overall elevation of ratings (cell means were 6.1655 and 6.3953 for the trained and untrained groups respectively) as anticipated in general hypothesis H_1. Contrary to the predictions in H_1, neither Participation

Table 8. Study One ANOVA Table--All Subjects

Source	df	SS	F^a	ω^2
Participation	1	1.6964	0.77	-
Training	1	31.4683	14.35*	.0013
Part x Train	1	0.4991	0.23	-
Subjects w. groups	4	8.7720	0.63	-
Categories	4	314.3566	22.57***	.0145
Part x Cat	4	27.6427	1.99	-
Train x Cat	4	7.5794	0.54	-
Part x Train x Cat	4	17.2175	1.24	-
Cat x Subj w. grp	16	82.7864	1.49	-
Professors	4	67.1426	4.82***	.0026
Part x Prof	4	47.9253	3.44**	.0016
Train x Prof	4	34.0416	2.44*	.0010
Part x Train x Prof	4	20.5515	1.48	-
Prof x Subj w. grp	16	54.7207	0.98	-
Cat x Prof	16	12071.2852	216.72***	.5786
Part x Cat x Prof	16	102.4909	1.84*	.0023
Train x Cat x Prof	16	79.9003	1.43	-
Part x Train x Cat x Prof	16	26.9005	0.48	-
Cat x Prof x Subj w. grp	64	167.8245	0.75	-
Residual	2183	7599.6374	-	-
Total	2382	20764.4391	-	-

[a]All effects were tested against Residual except for Participation, Training, and Part x Train, which were tested against Subjects w. groups.

*$p < .05$

**$p < .01$

***$p < .001$

Table 9. Study One ANOVA Table--Participant Subjects Only

Source	df	SS	F^a	ω^2
Training	1	11.6662	3.21	-
Subjects w. groups	2	7.2767	1.18	-
Categories	4	156.9696	5.37*	.0121
Train x Cat	4	9.8173	0.34	-
Cat x Subj w. grp	8	58.4766	2.36*	.0032
Professors	4	25.5242	2.06	-
Train x Prof	4	47.1598	3.81**	.0033
Prof x Subj w. grp	8	14.4242	0.58	-
Cat x Prof	16	6661.7041	134.54***	.6279
Train x Cat x Prof	16	57.5041	1.16	-
Cat x Prof x Subj w. grp	32	122.3188	1.24	-
Residual	1084	3354.7129	-	-
Total	1183	10527.5546	-	-

[a]All effects were tested against Residual except for Categories and Train x Cat, which were tested against Cat x Subj w. grp; and Training, which was tested against Subjects w. groups.

*$p < .05$

**$p < .01$

***$p < .001$

Table 10. Study One ANOVA Table--Non-participant Subjects Only

Source	df	SS	F^a	ω^2
Training	1	20.4132	27.30*	.0016
Subjects w. groups	2	1.4953	0.19	-
Categories	4	184.8966	11.97***	.0165
Train x Cat	4	15.4290	1.00	-
Cat x Subj w. grp	8	24.3098	0.79	-
Professors	4	89.2225	5.77***	.0072
Train x Prof	4	7.4333	0.48	-
Prof x Subj w. grp	8	40.2965	1.30	-
Cat x Prof	16	5511.2407	89.18***	.5322
Train x Cat x Prof	16	50.0515	0.81	-
Cat x Prof x Subj w. grp	32	45.5057	0.37	-
Residual	1099	4244.9245	-	-
Total	1198	10235.2187	-	-

[a]All effects were tested against Residual except for Training, which was tested against Subjects w. groups.

*$p < .05$

***$p < .001$

Table 11. Study One ANOVA Table--Trained Subjects Only

Source	df	SS	F[a]	ω^2
Participation	1	2.1316	3.69	-
Subjects w. groups	2	1.1542	0.17	-
Categories	4	179.2890	5.74*	.0148
Part x Cat	4	34.3782	1.10	-
Cat x Subj w. grp	8	62.4300	2.27*	.0035
Professors	4	33.1986	2.42*	.0019
Part x Prof	4	50.7707	3.70**	.0037
Prof x Subj w. grp	8	35.1496	1.28	-
Cat x Prof	16	5706.6370	103.97***	.5643
Part x Cat x Prof	16	56.1690	1.02	-
Cat x Prof x Subj w. grp	32	100.8442	0.92	-
Residual	1093	3749.4051	-	-
Total	1192	10011.5572	-	-

[a]All effects were tested against Residual except for Categories and Part x Cat, which were tested against Cat x Subj w. grp.; and Participation, which was tested against Subjects w. groups.

*$p < .05$

**$p < .01$

***$p < .001$

Table 12. Study One ANOVA Table--Untrained Subjects Only

Source	df	SS	F^a	ω^2
Participation	1	0.1454	0.04	-
Subjects w. groups	2	7.6178	1.08	-
Categories	4	142.5991	10.09***	.0120
Part x Cat	4	10.9008	0.77	-
Cat x Subj w. grp	8	20.3564	0.72	-
Professors	4	68.0058	4.81***	.0050
Part x Prof	4	17.3102	1.23	-
Prof x Subj w. grp	8	19.5712	0.69	-
Cat x Prof	16	6444.4720	114.03***	.5956
Part x Cat x Prof	16	73.2224	1.30	-
Cat x Prof x Subj w. grp	32	66.9803	0.59	-
Residual	1090	3850.2323	-	-
Total	1189	10721.4136	-	-

[a]All effects were tested against Residual except for Participation, which was tested against Subjects w. groups.

***$p < .001$

nor the Participation x Training interaction significantly influenced overall elevation.

The Category effect, representing variance attributable to the category being rated,[2] was statistically significant. Post-hoc analyses employing Duncan's multiple range test, with α set at .05 (see Kirk, 1968, pp. 93-94), indicated that Category D (Reasonableness of Workload, mean rating = 6.8391) ratings were significantly higher than those for Category C (Interest in Course and Material, mean rating = 6.4365) and Category E (Fairness of Testing and Grading, mean rating = 6.3579), which were in turn significantly higher than ratings for Category B (Ability to Present the Material, mean rating = 5.9252) and Category A (Relationships with Students, mean rating = 5.8450). Contrary to the predictions in general hypothesis H_3, neither Participation nor Training significantly reduced the Category effect, nor was there a significant Participation x Training x Category effect.

The Professor effect, representing variance attributable to the particular professor being rated,[3] was also statistically significant. The results of the Duncan multiple range test (with α again set at .05) showed that Professor P (mean rating = 5.9730) was rated significantly lower than were Professors N (mean rating = 6.2534), M (mean rating = 6.3400), L (mean rating = 6.3659), and O (mean rating = 6.4663). As

---

[2]As described in Chapter III, "true" mean scores across professors for all categories were set at or near 6.0, thus any variance attributable to Categories represents psychometric error.

[3]As described in Chapter III, "true" mean scores across categories for all professors were set at or near 6.0, thus any variance attributable to Professors also represents psychometric error.

predicted in general hypothesis H_2, Participation did significantly reduce the Professor effect (compare Tables 9 and 10), as did Training (compare Tables 11 and 12). Contrary to the H_2 prediction, the Participation x Training x Professor effect was not statistically significant.

The Category x Professor interaction effect was statistically significant, indicating that subjects were able to agree on assigning distinct behavioral profiles to professors (see Blumberg et al., 1966, p. 245). Such agreement is typically called "discriminant validity" (Friedman & Cornelius, 1976, p. 212). The findings further indicate that, as predicted in general hypothesis H_4, Participation significantly increased the Category x Professor effect (compare Tables 9 and 10). Contrary to the H_4 predictions, neither the Training x Category x Professor nor the Participation x Training x Category x Professor effects were statistically significant.

Discussion

The findings of Study One appear to support the following conclusions regarding the effects of the treatments on rating characteristics:

(1) Training significantly reduced the overall elevation of the ratings, whereas Participation did not. Given that the "true" mean rating was fixed at 6.0, this reduction in elevation can be interpreted as a reduction in leniency error (see Friedman & Cornelius, 1976, p. 212).

(2) Neither Participation nor Training significantly reduced the variance attributable to the category of behavior being evaluated, representing psychometric error.

(3) Both Participation and Training significantly reduced variance attributable to the professor being rated. Burnaska and Hollmann (1974, p. 307) would interpret this to mean that both treatments decreased the effects of consensual halo error in this set of ratings.

(4) Participation significantly increased the Category x Professor effect (discriminant validity; see Friedman & Cornelius, 1976, p. 212) while Training did not.

(5) There were no significant interactions among the treatments with regard to effects on any of the above characteristics of ratings. It appears that Participation and Training operate independently of each other, at least as far as these four characteristics of ratings are concerned.

The implications of these conclusions for subjective measurement of individual differences are considered in Chapter V. Note that the $\underline{\omega}^2$ statistics reported in Tables 8 through 12 suggest that while the effects of Participation on leniency error, and of both Participation and Training on halo error, may be statistically significant, their practical significance appears to be negligible. This does not seem to be the case, however, with the effects of Participation on discriminant validity. The fact that almost an additional ten percent of the variance in ratings may be accounted for in terms of discriminant validity when the subjects have participated in constructing the rating scales than when they have not (compare Tables 9 and 10) may have extremely important practical implications. The fact that almost 58

percent of the variance in ratings across all cells is accounted for by discriminant validity speaks well, of course, for the BARS themselves, and may explain why the other effects are so small.

Study Two

Purpose

The purpose of this study was to examine the main and interactive effects of Training and Participation on elevation of ratings *per category* of performance. General hypothesis H_5, stated in Chapter II, predicts the outcomes of this study. A MANOVA analysis, with each rater's mean rating (across professors) for each of the five categories serving as dependent variables, was performed to test the H_5 predictions. Individual ANOVAs for each category were carried out for interpretive purposes. The ω^2 statistic was calculated to determine the proportion of variance accounted for by each statistically significant effect.

Findings

The results of the MANOVA and five ANOVA evaluations are presented in Table 13. Cell means are found in Table 14. Although Participation resulted in a closer estimate of the "true" mean (6.0) for Category A (means were 6.0250 and 5.6642 for the Participant and Nonparticipant groups respectively), and Training significantly reduced elevation for Category B (means were 5.7360 and 6.1244 for the Trained and Untrained groups respectively, as compared to the 6.0 "true" mean), the results of the MANOVA tests were all non-significant. The only legitimate conclusion which can be drawn from this study is that neither Participation nor Training, nor their interaction, significantly

Table 13. Study Two MANOVA and ANOVA Tables

Source	df	SS	F	ω^2
		MANOVA[a]		
Participation	(5,88)	-	2.04	-
Training	(5,88)	-	1.59	-
Part x Train	(5,88)	-	1.06	-
		Category A: Relationships with Students		
Participation	1	3.1248	4.20*	.0312
Training	1	1.2834	1.73	-
Part x Train	1	2.8912	3.89	-
Residual	92	68.3707	-	-
Total	95	75.6702	-	-
		Category B: Ability to Present the Material		
Participation	1	0.4676	0.69	-
Training	1	3.6193	5.31*	.0434
Part x Train	1	0.1803	0.26	-
Residual	92	62.7527	-	-
Total	95	67.0198	-	-
		Category C: Interest in Course and Material		
Participation	1	1.0774	1.72	-
Training	1	0.0585	0.09	-
Part x Train	1	0.0776	0.12	-
Residual	92	57.7032	-	-
Total	95	58.9167	-	-

Table 13. Study Two MANOVA and ANOVA Tables (Cont'd)

Source	df	SS	F	ω^2
Category D: Reasonableness of Workload				
Participation	1	0.1218	0.13	-
Training	1	1.6017	1.65	-
Part x Train	1	0.4988	0.51	-
Residual	92	89.3657	-	-
Total	95	91.5881	-	-
Category E: Fairness of Testing and Grading				
Participation	1	1.1837	1.62	-
Training	1	1.0045	1.37	-
Part x Train	1	0.1001	0.14	-
Residual	92	67.3277	-	-
Total	95	69.6160	-	-

Note. The dependent variables in these analyses are the means of the subjects' ratings, across professors, for each category.

[a]The F tests for the MANOVA were approximated by the Hotelling-Lawley and Pillai traces.

*$p < .05$.

Table 14. Cell Means of Study Two Elevation Scores

Group	Category A	B	C	D	E
A. Both Participation and Training	6.0829	5.7096	6.4942	6.6763	6.1092
B. Participation Only	5.9671	6.0113	6.6004	7.0789	6.3783
C. Training Only	5.3750	5.7625	6.3392	6.7492	6.3958
D. Neither Participation nor Training	5.9533	6.2375	6.3317	6.8633	6.5358
"True" Elevation Scores	6.0000	6.0000	6.0000	5.9200	6.0200

Note. n = 24 for all elevation scores.

affects elevation of the ratings when examined on a category-by-category basis.

Discussion

General hypothesis H_5 predicted significant main and interactive effects of Participation and Training on elevation of ratings for each category; these predictions were not supported by this study. Apparently the effect of Training on elevation, found to be statistically significant in the Study One analysis of combined ratings, is not strong enough to be detected on a category-by-category basis. Implications of these findings are discussed in Chapter V.

Study Three

Purpose

Study Three was designed to examine the effects of Training, Participation, and the Training x Participation interaction on the dispersion (variance) of ratings per category. General hypothesis H_6 contains the specific predictions for this analysis. As in Study Two, a MANOVA analysis and five ANOVAs (one per category) were performed. The dependent variables were the variance (across professors) of each rater's ratings on each of the five categories. Again, ω^2 statistics were calculated.

Results

Since variance scores are squared deviates, there can be no scores below zero, thus the distribution may be non-normal due to a truncation at the lower end. One requirement of analysis of variance is that the distribution of scores under examination cannot deviate

greatly from normality (although ANOVA is quite robust in this respect; see Lindman, 1974, pp. 31-33). The variance scores across all categories were graphed in order to determine whether the assumption of normality was seriously violated (see Figure 3). The distribution is slightly skewed to the right (skewness = 0.3438) and flattened (kurtosis = -0.2960); however, these slight deviations from normality are well within the robustness limits of ANOVA (Lindman, 1974, pp. 31-33).

The MANOVA and ANOVA results are presented in Table 15. Although Participation led to significantly greater dispersion of ratings for Category E (variance scores were 11.2729 and 8.5346, respectively, for the Participant and Non-participant groups, as compared to the "true" variance of 8.0), the MANOVA indicates that neither Participation nor Training had statistically significant overall main effects on dispersion. The overall Participation x Training interaction effect was, however, statistically significant, as were the interaction effects on Categories B and C. The cell means of variance scores are presented in Table 16. Note that for Categories B and C, Training led to decreased dispersion when given to Participant subjects, but not to Non-participant subjects. Also note that for one category (C), the reduction in variance moved the obtained variance closer to the "true" variance, while for the other (B), it moved the obtained variance farther away from the "true" variance.

Discussion

Hypothesis H_6 predicted that both Participation and Training would significantly increase dispersion, thus counteracting the central tendency

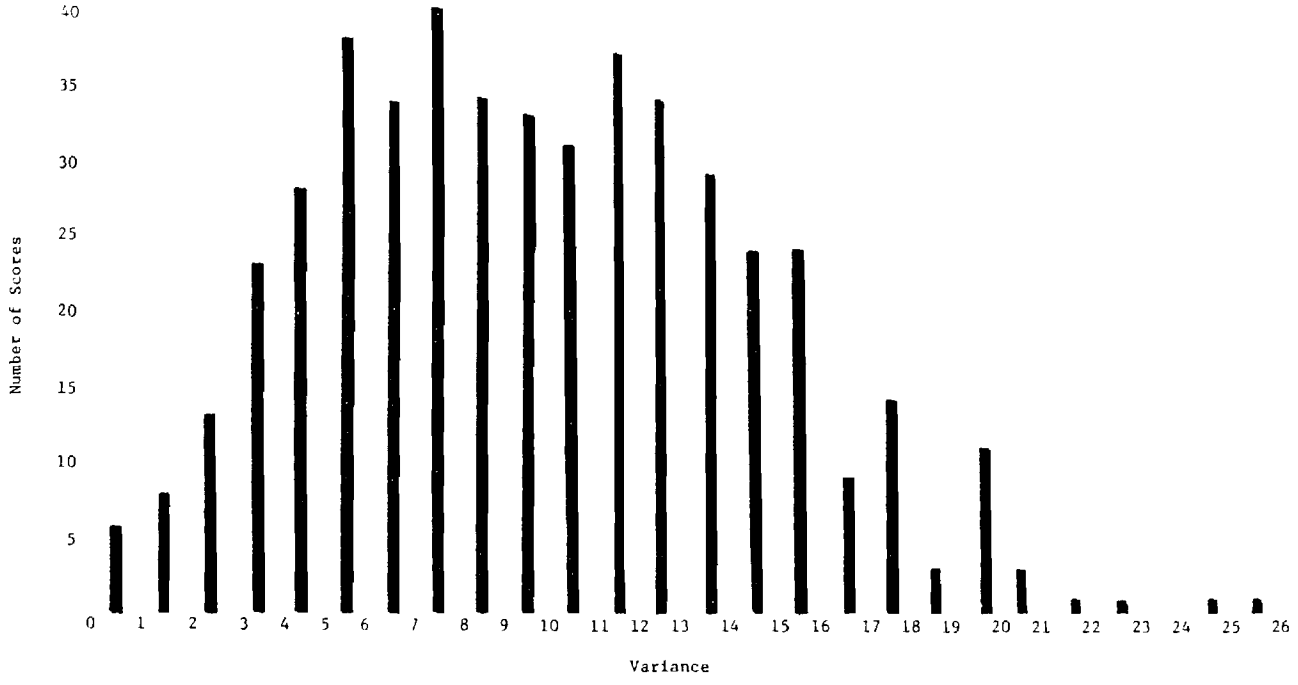


Figure 3. Distribution of Variance Scores Analyzed in Study Three

Table 15. Study Three MANOVA and ANOVA Tables

Source	df	SS	F	ω^2
		MANOVA[a]		
Participation	(5,88)	-	2.28	-
Training	(5,88)	-	0.95	-
Part x Train	(5,88)	-	4.21**	-
		Category A: Relationships with Students		
Participation	1	0.4013	0.02	-
Training	1	7.8404	0.39	-
Part x Train	1	0.2757	0.01	-
Residual	92	1859.1509	-	-
Total	95	1867.6684	-	-
		Category B: Ability to Present the Material		
Participation	1	5.7330	0.48	-
Training	1	43.6132	3.65	-
Part x Train	1	55.4070	4.64*	.0357
Residual	92	1099.4521	-	-
Total	95	1204.2054	-	-
		Category C: Interest in Course and Material		
Participation	1	2.0924	0.10	-
Training	1	52.6007	2.41	-
Part x Train	1	162.6276	7.45**	.0626
Residual	92	2009.6271	-	-
Total	95	2226.9479	-	-

Table 15. Study Three MANOVA and ANOVA Tables (Cont'd)

Source	df	SS	F	ω^2
Category D: Reasonableness of Workload				
Participation	1	0.3235	0.01	-
Training	1	20.9730	0.96	-
Part x Train	1	65.2724	3.00	-
Residual	92	2002.1393	-	-
Total	95	2088.7082	-	-
Category E: Fairness of Testing and Grading				
Participation	1	179.9496	8.14**	.0705
Training	1	0.0260	0.00	-
Part x Train	1	1.3976	0.06	-
Residual	92	2034.9205	-	-
Total	95	2216.2937	-	-

Note. The dependent variables in these analyses are the variances of the subjects' ratings, across professors, for each category.

[a]The F tests for the MANOVA were approximated by the Hotelling-Lawley and Pillai traces.

*p <.05

**p <.01

Table 16. Cell Means of Study Three Variance Scores

Group	Category A	B	C	D	E
A. Both Participation and Training	10.6887	5.5909	8.5461	10.8843	11.1687
B. Participation Only	11.1531	8.4584	12.6296	10.1699	11.3771
C. Training Only	10.4522	7.5991	10.8539	9.1190	8.6718
D. Neither Participation nor Training	11.1310	7.4277	9.7313	11.7030	8.3975
"True" Variance Scores	8.0000	8.0000	8.0000	7.3856	8.0016

Note. n = 24 for all variance scores.

error. The results of this study clearly refute these predictions--no overall difference in variance scores was produced by either treatment when considered as a main effect. The results of the interaction test are clearly quite surprising. There was no reason to expect Training to decrease dispersion, especially only when in combination with previous participation in the scale construction process. Perhaps the training program, following closely on the heels of the participation program, was perceived as an overemphasis on the error of extremism, and the subjects in Group A (Both Participation and Training) attempted to counteract extremism with less dispersion among ratings. The fact that the interaction effect accounted for only 3.27 percent of the variance in Category B scores and 6.26 percent in Category C scores certainly diminishes the impact of this finding. Nonetheless, the question of a possible Participation-Training "overkill" effect appears worthy of future investigation.

Study Four

Purpose

The purpose of Study Four was to examine the effects of Training, Participation, and their combination on mean intercorrelation among category ratings, a statistic commonly employed to operationally define halo error. General hypothesis H_7 expressed the expected findings of this study. For each ratee within each cell of the design, intercorrelations of category ratings (across subjects, such that the n for each correlation was approximately 24) were calculated and converted to Z scores through the use of Pearson's r-to-Z transformation (see Guilford

& Fruchter, 1978, p. 522). The mean intercorrelation for each ratee within each cell was determined by calculating the mean Z value, then reconverting to r. These data are displayed in Table 17[4].

Findings

The results of the five χ^2 analyses of equivalence are presented in Table 18. These results indicate that for every category, neither Participation, Training, nor their combination resulted in measures of the halo error which were significantly smaller than those obtained from the control group.

Discussion

General hypothesis H_7, which predicted significant reductions in intercorrelations among category ratings to result from Training and Participation, was not supported by the results of this study. When examined in this manner, neither Participation nor Training seems effective in reducing halo error. Of course, the fact that no evidence of halo error was detected for the control group mitigates the importance of this finding. Implications of the results of this study are considered in the next chapter.

Study Five

Purpose

The purpose of Study Five was to examine the effects of Training, Participation, and the Training x Participation interaction on measures

---

[4]Note that none of the mean correlation coefficients were statistically significant, indicating virtual absence of the halo error in this set of ratings.

Table 17. Mean Category Intercorrelations for Each Simulated Professor Within Each Cell of the Design

Group	Simulated Professor				
	L	M	N	O	P
A. Both Participation and Training	.045 (.046)	.005 (.009)	.150 (.153)	.010 (.012)	-.095 (-.095)
B. Participation Only	.160 (.165)	-.035 (-.036)	.105 (.108)	.090 (.090)	.005 (.005)
C. Training Only	.250 (.249)	.215 (.223)	.075 (.079)	.180 (.185)	-.010 (-.011)
D. Neither Participation nor Training	.235 (.242)	.155 (.158)	.075 (.079)	.140 (.144)	.030 (.034)

Note. The number in parentheses below each correlation coefficient is its corresponding approximate Z value. None of the mean correlation coefficients are statistically significant at $\alpha = .05$. $N = 24$ for all mean correlation coefficients.

Table 18. Summary of Study Four χ^2 Tests

Simulated Professor	df	χ^2
L	3	0.55
M	3	0.94
N	3	0.08
O	3	0.35
P	3	0.19

Note. None of the χ^2 tests were significant at $\alpha = .05$.

of the reliability and validity of the obtained ratings. Four such measures were calculated for each category within each cell of the experimental design: (a) Ebel's (1951) intraclass reliability coefficient; (b) Ebel's (1951) one-rater reliability coefficient; (c) the intraclass correlation coefficient (Guilford, 1954, pp. 395-397) between group mean rating and "true" score, a measure of validity; and (d) the corresponding one-rater validity coefficient (Guilford, 1954, p. 407). These data are presented in Table 19. Since correlation coefficients cannot be assumed to be normally distributed, all rs were transformed to Z scores for use in further analyses (see Guilford & Fruchter, 1978, p. 522). The four general hypotheses under consideration in this study are presented in Chapter II. A separate ANOVA was performed for each dependent variable. Categories served as a blocking variable in these analyses. The ω^2 statistic was calculated for each statistically significant effect.

Findings

The ANOVA results for all four dependent variables are found in Table 20. With regard to intraclass reliability scores, Participation significantly increased reliability (Z scores for the Participant and Non-participant groups were 2.2699 and 2.1068 respectively, corresponding to r scores of .975 and .970). There was also a statistically significant Category effect: The Duncan Multiple range test, with $\alpha = .05$, shows that Categories A ($Z = 2.3060$, $r = .980$), D ($Z = 2.2778$, $r = .975$), C ($Z = 2.2545$, $r = .975$), and E ($Z = 2.1948$, $r = .975$) were all rated significantly more reliably than was Category B ($Z = 1.9088$, $r = .955$). With respect to one-rater reliability scores, Participation was again

Table 19. Study Five Reliability and Validity Scores

Category[a]	Intra-Class Reliability	One-Rater Reliability	Intra-Class Validity	One-Rater Validity
	Group A: Both Participation and Training			
A	.9824	.7012	.9878	.8345
B	.9674	.5735	.7830	.5930
C	.9775	.6463	.7522	.6115
D	.9841	.7221	.9289	.7958
E	.9860	.7491	.9753	.8501
Average[b]	.9750	.6800	.9250	.7500
	Group B: Participation Only			
A	.9865	.7547	.9742	.8519
B	.9706	.5811	.8124	.6286
C	.9852	.7383	.8279	.7167
D	.9754	.6286	.9506	.7633
E	.9840	.7245	.9608	.8244
Average[b]	.9750	.6800	.9250	.7600
	Group C: Training Only			
A	.9804	.6763	.9651	.8013
B	.9393	.3940	.7293	.4724
C	.9711	.5837	.7658	.5935
D	.9788	.6580	.9609	.7878
E	.9622	.5146	.8167	.5973
Average[b]	.9700	.5700	.8850	.6650

Table 19. Study Five Reliability and Validity Scores (Cont'd)

Category[a]	Intra-Class Reliability	One-Rater Reliability	Intra-Class Validity	One-Rater Validity
Group D: Neither Participation nor Training				
A	.9796	.6671	.9599	.7919
B	.9532	.4589	.7604	.5277
C	.9843	.7226	.7602	.6516
D	.9759	.6278	.9662	.7750
E	.9717	.5890	.8627	.6714
Average[b]	.9700	.6150	.8900	.6900

[a]Categories are as follows:
A. Relationships with Students
B. Ability to Present the Material
C. Interest in Course and Material
D. Reasonableness of Workload
E. Fairness of Testing and Grading

[b]Averages were calculated by converting r to Z, averaging Z scores, then reconverting Z to r (see Guilford & Fruchter, 1978, p. 522).

Table 20. Study Five ANOVA Tables

Source	df	SS	F	ω^2
Intra-Class Reliability Scores				
Participation	1	0.1330	12.10*	.1495
Training	1	0.0057	0.52	-
Category	4	0.4177	9.50*	.4580
Part x Train	1	0.0006	0.05	-
Part x Cat	4	0.1098	2.50	-
Train x Cat	4	0.0941	2.14	-
Residual	4	0.0440	-	-
Total	19	0.8049	-	-
One-Rater Reliability Scores				
Participation	1	0.1160	38.38**	.2363
Training	1	0.0091	3.02	-
Category	4	0.2340	19.36**	.4641
Part x Train	1	0.0045	1.48	-
Part x Cat	4	0.0473	3.91	-
Train x Cat	4	0.0523	4.33	-
Residual	4	0.0121	-	-
Total	19	0.4753	-	-
Intra-Class Validity Scores				
Participation	1	0.2376	16.46*	.0453
Training	1	0.0000	0.00	-
Category	4	3.8591	66.81***	.7713
Part x Train	1	0.0046	0.32	-
Part x Cat	4	0.6650	11.51*	.1232
Train x Cat	4	0.0899	1.56	-
Residual	4	0.0578	-	-
Total	19	4.9140	-	-

Table 20. Study Five ANOVA Tables (Cont'd)

Source	df	SS	F	ω^2
		One-Rater Validity Scores		
Participation	1	0.1365	31.34**	.1266
Training	1	0.0067	1.54	-
Category	4	0.7329	42.08**	.6853
Part x Train	1	0.0007	0.17	-
Part x Cat	4	0.1237	7.10*	.0233
Train x Cat	4	0.0215	1.24	-
Residual	4	0.0174	-	-
Total	19	1.0394	-	-

* $p < .05$

** $p < .01$

*** $p < .001$

found to significantly increase reliability ($\underline{Z}$ scores for the Participant and Non-participant groups were 0.8368 and 0.6845 respectively, corresponding to $\underline{r}$ scores of .680 and .590). Again, there was also a statistically significant Category effect: The Duncan multiple range test, with $\underline{\alpha}$ = .05, indicates that Categories A ($\underline{Z}$ = 0.8655, $\underline{r}$ = .695), C ($\underline{Z}$ = 0.8193, $\underline{r}$ = .670), D ($\underline{Z}$ = 0.7895, $\underline{r}$ = .655), and E ($\underline{Z}$ = 0.7758, $\underline{r}$ = .650) were all rated significantly more reliably than was Category B ($\underline{Z}$ = 0.5533, $\underline{r}$ = .500).

The analyses of the validity data continue to support the value of Participation, and to document significant inter-category differences. The intraclass validity ANOVA indicates that Participation leads to a significant increase in validity ($\underline{Z}$ scores for the Participant and Non-participant groups were 1.6393 and 1.4213 respectively, corresponding to $\underline{r}$ scores of .925 and .885). The Category groupings, as determined by Duncan's multiple range test, with $\underline{\alpha}$ = .05, were as follows: Ratings for Category A ($\underline{Z}$ = 2.1088, $\underline{r}$ = .970) were significantly more valid than those for Categories D ($\underline{Z}$ = 1.8538, $\underline{r}$ = .950) and E ($\underline{Z}$ = 1.6415, $\underline{r}$ = .925), which were in turn significantly more valid than those for Categories C ($\underline{Z}$ = 1.0260, $\underline{r}$ = .770) and B ($\underline{Z}$ = 1.0215, $\underline{r}$ = .770). Similar findings were obtained for the one-rater validity ANOVA. Participation significantly increased one-rater validity scores ($\underline{Z}$ scores for the Participant and Non-Participant groups were 0.9955 and 0.8303 respectively, corresponding to $\underline{r}$ scores of .755 and .680). Duncan's multiple range test, with $\underline{\alpha}$ = .05, indicates that the Category groups in terms of one-rater validity scores are as follows: Ratings for Category A ($\underline{Z}$ = 1.1535, $\underline{r}$ = .815) were

significantly more valid than those for all other categories except D ($\underline{Z}$ = 1.0430, $\underline{r}$ = .775). Ratings for Category D were significantly more valid than those for all remaining Categories save E ($\underline{Z}$ = 0.9773, $\underline{r}$ = .750). Those for Category E were significantly more valid than those for Category C ($\underline{Z}$ = 0.7648, $\underline{r}$ = .640), which were in turn significantly more valid than those for Category B ($\underline{Z}$ = 0.6260, $\underline{r}$ = .555).

In both validity ANOVAs, a significant Participation x Category interaction effect was found. Examinations of the cell means (see Table 21) indicate that Participation increased validity for all Categories except D (Reasonableness of Workload). With respect to intraclass validity, Non-participant Category D scores appear to be higher than Participant Category D scores; whereas for one-rater validity, the Category D scores appear roughly equal for the two groups.

Discussion

General hypotheses H_8, H_9, H_{10}, and H_{11}, presented in Chapter II, predict that Participation and Training will have significant main and interactive effects on intraclass reliability, one-rater reliability, intraclass validity, and one-rater validity of the performance appraisal ratings of the five simulated professors. The predictions regarding Training and the Training x Participation interaction were not confirmed. Apparently the training program's effects on the leniency and consensual halo errors (see Study One) were not manifested in terms of improvements in either the reliability or the validity of the ratings.

The positive effects of Participation, however, are strongly documented by the results of this study. Despite the fact that these results

Table 21. Study Five Participation x Category Cell Means for Intra-Class and One-Rater Validity Scores

Participation	Category	Intra-Class Validity		One-Rater Validity	
		Z	r	Z	r
Yes	A	2.2675	.975	1.2220	.835
Yes	B	1.0860	.795	0.7055	.605
Yes	C	1.0500	.780	0.8030	.665
Yes	D	1.7275	.935	1.0405	.775
Yes	E	2.0655	.965	1.2065	.835
No	A	1.9500	.960	1.0850	.795
No	B	0.9570	.740	0.5465	.495
No	C	1.0020	.760	0.7265	.620
No	D	1.9800	.960	1.0455	.780
No	E	1.2175	.835	0.7480	.962

were obtained in a controlled experiment using simulated ratees, the findings that rater participation in scale construction accounted for 14.95 percent of the variance in intraclass reliability scores, 23.63 percent of the variance in one-rater reliability scores, 4.53 percent of the variance in intraclass validity scores, and 12.66 percent of the variance in one-rater validity scores is encouraging, and suggests that the application of participative techniques in rating scale construction is worthy of further investigation. Additional implications of the findings of Study Five are presented in Chapter V.

The findings regarding the Category and Participation x Category effects come as no great surprise. Several of the articles cited in Table 1 section VII (Aspects of the Behavioral Characteristic) suggest that for one reason or another some behavioral characteristics are easier to rate than are others (Ferguson, 1949a; Stockford & Bissell, 1949). It appeared during the scale construction process that Categories A, D, and E were more distinct and more easily understood by the subjects than were Categories B and C. These latter two categories did not elicit as many clearly representative critical incidents in Step (2) of the scale construction process as did the other three categories. In Step (4) of the BARS development process, subjects sorted incidents into categories. The fact that the subjects often disagreed in the placement of incidents into Categories B and C, yet had relatively little difficulty in placing incidents into the other three categories, further documents this lack of clarity of understanding of Categories B and C. The high positive correlations between ratings of Categories B and C reported in

Table 17 also support this conclusion. Relationships with Students, Reasonableness of the Workload, and Fairness of Testing and Grading may be more distinct, more easily observed categories of behavior than Ability to Present the Material, which can be interpreted quite broadly, and Interest in Course and Material, which probably calls for greater inference on the subjects' part than do the other categories. Ratings for the less clear categories may be more susceptible to halo and logical errors (Symonds, 1925). This suggests, of course, that the BARS approach to performance appraisal may not be equally appropriate or effective for all categories of behavior, or at least might require special care in category definition. Research with additional behavioral categories, subject pools, and settings may be useful in examining this issue.

Study Six

Purpose

Study Six was designed to measure the effects of Training, Participation, and the Training x Participation interaction on Attitudes. General hypothesis H_{12}, stated in Chapter II, presents the predicted outcomes of the study. The analysis was an ANOCOV, with Post-treatment Attitude scores (Form B) serving as the dependent variable and Pre-treatment Attitude scores (Form A) as the covariate variable.

Findings

The outcomes of the ANOCOV are presented in Table 22. As expected, Pre-treatment Attitude scores significantly related to Post-treatment Attitude scores, thus the ANOCOV design was certainly appropriate. The effects of Participation on Post-treatment Attitude were statistically

Table 22. Study Six ANOCOV Table

Source	df	SS	F	ω^2
Covariance (Pre-treatment Attitude)	1	11.8588	38.03***	.2566
Participation	1	4.2156	13.52***	.0867
Training	1	0.0044	0.01	-
Part x Train	1	0.2391	0.77	-
Residual	91	28.3754	-	-
Total	95	44.6933	-	-

Note. Dependent variable = Post-treatment Attitude score.

*** $p < .001$

significant, but were in the opposite direction from that hypothesized: Adjusted mean Post-treatment Attitude scores were 6.60 for the Participant group and 7.02 for the Non-participant group. (Compare these with the "normative" mean Form B scores of 7.05 and 7.11 reported in Chapter III.) Neither Training nor the Participation x Training interaction significantly affected Post-treatment Attitude.

Discussion

General hypothesis H_{12}, formulated on the basis of several empirical and theoretical investigations of Participation and Training, predicted that the two treatments would improve subjects' attitudes toward the performance appraisal rating process. These predictions are clearly unsupported by the data. Although Participation did significantly affect Post-treatment Attitude scores, and accounted for 8.67 percent of the variance in those scores, the result was a less positive attitude toward the performance appraisal process. One possible explanation for this somewhat startling finding could be that the attitude measurement instrument is inadequate in terms of construct validity. However, the data presented in Chapter III suggest otherwise.

A second possible explanation is that the treatments "backfired" in their attempts to improve attitudes. It is possible that the student subjects in the Participant group felt over-exposed to material concerning performance appraisal, and their Post-treatment Attitude scores reflect their impatience with the experimental procedures. However, if this were the case, one might also expect a statistically significant Training effect, since the training program consisted of two hours of

exposure to material relevant to performance appraisal. A statistically significant Participation x Training interaction effect could certainly be expected, since subjects in Group A, who received both treatments for a total exposure of six hours to performance-appraisal relevant material, would be expected to be least pleased with the experiment if exposure to such material were deemed unpleasant. Subjects in other groups, having received less exposure, would not be expected to express as much disfavor with the performance appraisal process. The data do not appear to support this second explanation. In fact, the Participant subjects who were not trained had a slightly lower mean adjusted Post-treatment Attitude score (6.56) than those who were also exposed to the training program (6.64). The fact that the Attitude scores for all groups were above neutral (6.0) also detracts from this explanation, since subjects expressing displeasure would be expected to report scores in the negative direction.

A third possible explanation exists. Several authors (Bavelas & Strauss, 1961, p. 590; Lowin, 1968, p. 80; Strauss, 1963, p. 70; Tannenbaum, 1966, p. 101) have suggested that Participant subjects whose ideas, suggestions, and products are not implemented may become disenchanted with whatever procedures are subsequently or currently employed. Subjects in this experiment were clearly informed that the BARS resulting from their efforts were for experimental use only and would not be put into operational use in the foreseeable future. Perhaps after seeing their own scales, and feeling the pride of accomplishment and ownership which numerous authors (e.g., Lowin, 1968; Maier, 1967; Tannenbaum, 1966;

Vroom & Yetton, 1973; Wood, 1973) suggest results from participative decision making, the Participant subjects were less enthusiastic about current teacher performance appraisal practices. Although the Attitude scales were designed to measure attitude toward performance appraisal in the abstract, it is quite possible that subjects were responding with reference to the concrete teacher evaluation schemes to which they had been exposed during their education. A student committed to the concept of teacher evaluation, yet disenchanted with extant evaluation devices, might be expected to express a slightly positive overall Attitude score of the magnitude found in this study.

These three possible explanations, and others which may not have been mentioned, must stand as speculation until new research resolves this question. Of the three, the third appears most likely given the evidence available. The third explanation is also the one supported by data collected through informal interviews with several of the Participant subjects. The gist of their commentary was, "After building what we felt was a really fine teacher evaluation scale, we just weren't happy with the scales we are now using." Perhaps had attitude toward the new scale been measured, the third explanation may have been strongly supported.

Study Seven

Purpose

Study Seven was intended to examine the main and interactive effects of Participation and Training on Knowledge. General hypothesis H_{13} predicts that Training will lead to increased Knowledge, whereas

Participation and Participation x Training will not affect Knowledge. The analysis was an ANOCOV, with Post-treatment Knowledge scores serving as the dependent variable and Pre-treatment Knowledge scores as the covariate variable. A second ANOCOV, with Post-treatment "Irrelevant" Knowledge scores serving as the dependent variable and Pre-treatment "Irrelevant" Knowledge scores as the covariate variable, was performed to examine any possible Hawthorne effects of the treatments.

Findings

The results of the two ANOCOVs are presented in Table 23. Again, the covariate variables accounted for statistically and practically significant proportions of the variance of the two dependent variables, supporting the use of ANOCOV. The findings support the H_{13} predictions that Training would increase Knowledge (adjusted cell means for the Trained and Untrained groups were 43.46 and 40.58 respectively; compare with "normative" scores reported in Chapter III ranging from 38.49 to 40.62), while Participation would not. Contrary to the H_{13} prediction, the Participation x Training interaction effect on Knowledge was statistically significant. An examination of the adjusted cell means (see Table 24) indicates that Training was more effective in enhancing Knowledge when subjects had already participated in constructing the BARS. The results of the ANOCOV of "Irrelevant" Knowledge scores was encouraging: Apparently neither Participation nor Training produced changes in an unintended variable.

Table 23. Study Seven ANOCOV Tables

Source	df	SS	$\underline{F}$	$\underline{\omega}^2$
Post-treatment Knowledge Score				
Covariance (Pre-treatment Knowledge)	1	534.0683	50.37***	.2915
Participation	1	5.0317	0.47	-
Training	1	208.5579	19.67***	.1102
Part x Train	1	72.3516	6.82*	.0344
Residual	91	964.8969	-	-
Total	95	1784.9063	-	-
Post-treatment "Irrelevant" Knowledge Score				
Covariance (Pre-treat. "Irrel." Know.)	1	968.4363	58.80***	.3726
Participation	1	39.8605	2.42	-
Training	1	1.3949	0.08	-
Part x Train	1	13.3567	0.81	-
Residual	91	1498.6912	-	-
Total	95	2521.7396	-	-

* $\underline{p} < .05$

*** $\underline{p} < .001$

Table 24. Study Seven Adjusted Cell Means

Group	Adjusted Mean Post-treatment Knowledge Score[a]
A. Both Participation and Training	44.0961
B. Participation Only	39.3838
C. Training Only	42.8174
D. Neither Participation nor Training	41.5778

[a]Adjusted for Pre-treatment Knowledge score.

Discussion

The findings of Study Six validate the belief that a training program such as that employed in this study can increase subjects' knowledge of the rating process, and support the findings of many research projects involving rater training reviewed in Chapter I. They also tend to strengthen claims for the construct validity of the Knowledge Scale, especially in view of the fact that Training did not lead to an increase in "Irrelevant" Knowledge. The Participation x Training interaction can most likely be explained in terms of increased motivation on the part of subjects who had participated in the construction of the BARS which they were being trained to use. Additional research on the motivating properties of Participation, as requested in the discussion of the results of the previous study, might shed new light on the interpretation of this interaction effect.

Study Eight

Purpose

The model suggests that the effects of Training on characteristics of ratings can be accounted for in terms of its effects on the mediating variables. Study Eight was carried out in order to test this prediction with regard to overall elevation of the ratings. The specific statistical predictions for this study are presented in general hypothesis H_{14}, found in Chapter II. Although the prediction regarding the effects of Participation is moot, since Study One found no significant effects of Participation on overall elevation of the ratings, Participation and the Participation x Training interaction were included in the analyses in

order to explore for possible unexpected effects of the mediating variables. Also, since Training was found to significantly affect only Knowledge, the use of Attitude as a covariate variable is not necessary; yet Attitude was also retained in the analyses for similar exploratory purposes.

Four analyses were carried out in this study. The first, a factorial ANOVA with Participation and Training as factors and ratings as the dependent variable, employing the Subjects within Groups Mean Square as the error term for all F tests, corresponds to the "Between Subjects" portion of the overall Study One analysis. Three ANOCOVs were also carried out. They differed from the first analysis only in terms of the covariate variables employed; Change in Attitude in the first, Change in Knowledge in the second, and Changes in Attitude and Knowledge in the third. The Change scores were constructed by taking each subject's algebraic difference between Pre- and Post-treatment scores for each scale.[5]

Findings

The results of the four analyses are presented in Table 25. The data suggest that the effects of Training on general elevation of the ratings are not accounted for by covarying change scores in Attitude, Knowledge, or their combination. The exploratory analyses did not uncover

---

[5]Change in Attitude and Change in Knowledge were correlated .025. This correlation coefficient, calculated on the basis of 96 pairs of scores, is nonsignificant (Guilford, 1965, pp. 580-581).

Table 25. Study Eight ANOVA and ANOCOV Tables

Source	df	SS	F[a]	ω^2
No Covariates				
Participation	1	1.6658	0.69	-
Training	1	31.4989	13.19*	.0014
Part x Train	1	0.5805	0.24	-
Subj. w. groups	4	9.5553	-	-
Residual	2375	20721.1386	-	-
Total	2382	20764.4391	-	-
Covarying Change in Attitude				
Covariance (Change in Attitude)	1	3.7470	2.06	-
Participation	1	0.5831	0.32	-
Training	1	35.0698	19.29*	.0016
Part x Train	1	0.5956	0.33	-
Subj. w. groups	4	7.2727	-	-
Residual	2374	20717.1709	-	-
Total	2382	20764.4391	-	-
Covarying Change in Knowledge				
Covariance (Change in Knowledge)	1	5.0135	3.31	-
Participation	1	2.1118	1.39	-
Training	1	44.9213	29.64**	.0021
Part x Train	1	0.0589	0.04	-
Subj. w. groups	4	6.0620	-	-
Residual	2374	20706.2715	-	-
Total	2382	20764.4391	-	-

Table 25. Study Eight ANOVA and ANOCOV Tables (Cont'd)

Source	df	SS	F^a	ω^2
Covarying Changes in Attitude and Knowledge				
Covariance				
(Change in Attitude)	1	4.0315	3.78	-
(Change in Knowledge)	1	5.1035	4.70	-
Participation	1	0.8356	0.78	-
Training	1	49.8765	46.75**	.0024
Part x Train	1	0.0745	0.07	-
Subj. w. groups	4	4.2674	-	-
Residual	2373	20700.3399	-	-
Total	2382	20764.4391	-	-

[a]All F tests employed $MS_{subj.\ w.\ groups}$ as the error term.

*$p < .05$

**$p < .01$

any irregularities from the predicted relationships. Note that the covariate effects were nonsignificant in all analyses.

Discussion

Covarying changes in Attitude and Knowledge scores produced negligible increases in the proportion of variance in general elevation (leniency) accounted for by Training; certainly the model's prediction that the statistically significant effects of Training on leniency would be "washed out" by covarying change in Knowledge due to Training was not supported. This suggests that some additional variable or set of variables mediates the effects of Training on at least this one characteristic of ratings. Implications of the findings of this study for the hypothetical model are considered in more detail in Chapter V.

Study Nine

The stated purpose of Study Nine was to attempt to explain any statistically significant effects of Training, Participation, or the Training x Participation interaction on elevation of ratings per category, found in Study Two, in terms of the two hypothesized mediating variables, Attitude and Knowledge. Since the MANOVA results reported in Study Two uncovered no such statistically significant effects, there was nothing to explain--the study was moot and was not carried out.[6]

---

[6]The MANOCOVs and supporting ANOCOVs were, however, performed for exploratory purposes. The results did not differ from those of Study Two when Change in Attitude, Change in Knowledge, or Changes in Attitude and Knowledge were held constant through the use of analysis of covariance. Hypothesis H_{15} was, of course, automatically supported. None of the covariance effects were statistically significant.

Study Ten

Purpose

The purpose of Study Ten was similar to that of Study Nine: To attempt to explain any significant effects found in Study Three in terms of the mediating variables found to be significantly affected by the treatments in Studies Six and Seven. Recall that Study Three found a significant Participation x Training interaction effect on dispersion of the ratings when examined on a category-by-category basis. Furthermore, the individual ANOVAs for Categories B and C indicated that Training decreased dispersion for these categories when in combination with Participation, but not otherwise. General hypothesis H_{16} suggests that these effects will "wash out" when changes in Attitude and Knowledge scores due to the effects of the treatments are held constant through the use of analysis of covariance. Study Ten was intended to test these predictions. Three MANOCOV-ANOCOV sets were performed, with Change in Attitude covaried in the first set, Change in Knowledge in the second, and Changes in Attitude and Knowledge in the third. The dependent variables in these analyses were the variances of the subjects' ratings, across professors, for each category.

Findings

The results of the three MANOCOV-ANOCOV sets are presented in Table 26. It is readily apparent that the H_{16} predictions were not supported by the data. Covarying Change in Knowledge produced no differences from the results reported in Study Three, for either the multivariate or univariate tests. Covarying Change in Attitudes, either separately

Table 26. Study Ten MANOCOV and ANOCOV Tables

Source	df	SS	F	ω^2
MANOCOV--Covarying Change in Attitude[a]				
Participation	(5,87)	-	2.43*	-
Training	(5,87)	-	0.97	-
Part x Train	(5,87)	-	4.18**	-
Category A ANOCOV--Covarying Change in Attitude				
Covariance (Change in Attitude)	1	0.1401	0.01	-
Participation	1	0.5997	0.03	-
Training	1	7.5256	0.37	-
Part x Train	1	0.2757	0.01	-
Residual	91	1859.1273	-	-
Total	95	1867.6684	-	-
Category B ANOCOV--Covarying Change in Attitude				
Covariance (Change in Attitude)	1	6.5604	0.54	-
Participation	1	2.9897	0.25	-
Training	1	40.7008	3.37	-
Part x Train	1	55.4070	4.59*	.0356
Residual	91	1098.5473	-	-
Total	95	1204.2054	-	-

Table 26. Study Ten MANOCOV and ANOCOV Tables (Cont'd)

Source	df	SS	F	ω^2
Category C ANOCOV--Covarying Change in Attitude				
Covariance (Change in Attitude)	1	0.0479	0.00	-
Participation	1	2.0874	0.09	-
Training	1	53.3991	2.42	-
Part x Train	1	162.6276	7.37**	.0625
Residual	91	2008.7859	-	-
Total	95	2226.9479	-	-
Category D ANOCOV--Covarying Change in Attitude				
Covariance (Change in Attitude)	1	8.6346	0.39	-
Participation	1	2.1881	0.10	-
Training	1	17.0666	0.78	-
Part x Train	1	65.2724	2.98	-
Residual	91	1995.5466	-	-
Total	95	2088.7082	-	-
Category E ANOCOV--Covarying Change in Attitude				
Covariance (Change in Attitude)	1	1.4993	0.07	-
Participation	1	206.7505	9.38**	.0825
Training	1	0.9792	0.04	-
Part x Train	1	1.3976	0.06	-
Residual	91	2005.6671	-	-
Total	95	2216.2937	-	-

Table 26. Study Ten MANOCOV and ANOCOV Tables (Cont'd)

Source	df	SS	F	ω^2
MANOCOV--Covarying Change in Knowledge[a]				
Participation	(5,87)	-	2.23	-
Training	(5,87)	-	0.89	-
Part x Train	(5,87)	-	4.10**	-
Category A ANOCOV--Covarying Change in Knowledge				
Covariance (Change in Knowledge)	1	18.7876	0.93	-
Participation	1	0.1124	0.01	-
Training	1	2.2687	0.11	-
Part x Train	1	1.9314	0.10	-
Residual	91	1844.5682	-	-
Total	95	1867.6684	-	-
Category B ANOCOV--Covarying Change in Knowledge				
Covariance (Change in Knowledge)	1	1.2820	0.11	-
Participation	1	6.1424	0.51	-
Training	1	42.6471	3.55	-
Part x Train	1	61.6258	5.13*	.0408
Residual	91	1092.5081	-	-
Total	95	1204.2054	-	-

Table 26. Study Ten MANOCOV and ANOCOV Tables (Cont'd)

Source	df	SS	F	ω^2
Category C ANOCOV--Covarying Change in Knowledge				
Covariance (Change in Knowledge)	1	65.8955	3.00	-
Participation	1	0.7898	0.04	-
Training	1	24.4669	1.11	-
Part x Train	1	136.3399	6.21*	.0506
Residual	91	1999.4559	-	-
Total	95	2226.9479	-	-
Category D ANOCOV--Covarying Change in Knowledge				
Covariance (Change in Knowledge)	1	6.9062	0.32	-
Participation	1	0.1509	0.01	-
Training	1	15.6697	0.72	-
Part x Train	1	73.6970	3.37	-
Residual	91	1992.2845	-	-
Total	95	2088.7082	-	-
Category E ANOCOV--Covarying Change in Knowledge				
Covariance (Change in Knowledge)	1	12.8465	0.58	-
Participation	1	174.2064	7.82**	.0679
Training	1	1.1396	0.05	-
Part x Train	1	0.2884	0.01	-
Residual	91	2027.8128	-	-
Total	95	2216.2937	-	-

Table 26. Study Ten MANOCOV and ANOCOV Tables (Cont'd)

Source	df	SS	$\underline{F}$	$\underline{\omega}^2$
MANOCOV--Covarying Changes in Attitude and Knowledge[a]				
Participation	(5,86)	-	2.39*	-
Training	(5,86)	-	0.90	-
Part x Train	(5,86)	-	4.08**	-
Category A ANOCOV--Covarying Changes in Attitude and Knowledge				
Covariance				
(Change in Attitude)	1	0.2326	0.01	-
(Change in Knowledge)	1	18.7876	0.92	-
Participation	1	0.2449	0.01	-
Training	1	2.0351	0.10	-
Part x Train	1	1.9456	0.09	-
Residual	90	1844.4227	-	-
Total	95	1867.6684	-	-
Category B ANOCOV--Covarying Changes in Attitude and Knowledge				
Covariance				
(Change in Attitude)	1	6.7096	0.55	-
(Change in Knowledge)	1	1.2820	0.11	-
Participation	1	3.2765	0.27	-
Training	1	39.6195	3.27	-
Part x Train	1	61.4431	5.06*	.0405
Residual	90	1091.8747	-	-
Total	95	1204.2054	-	-

Table 26. Study Ten MANOCOV and ANOCOV Tables (Cont'd)

Source	df	SS	F	ω^2
Category C ANOCOV--Covarying Changes in Attitude and Knowledge				
Covariance				
(Change in Attitude)	1	0.0003	0.00	-
(Change in Knowledge)	1	65.8955	2.97	-
Participation	1	0.8514	0.04	-
Training	1	24.7284	1.11	-
Part x Train	1	136.5476	6.15*	.0508
Residual	90	1998.9247	-	-
Total	95	2226.9479	-	-
Category D ANOCOV--Covarying Changes in Attitude and Knowledge				
Covariance				
(Change in Attitude)	1	9.0285	0.41	-
(Change in Knowledge)	1	6.9062	0.31	-
Participation	1	1.7095	0.08	-
Training	1	12.0588	0.55	-
Part x Train	1	74.3339	3.37	-
Residual	90	1984.6713	-	-
Total	95	2088.7082	-	-
Category E ANOCOV--Covarying Changes in Attitude and Knowledge				
Covariance				
(Change in Attitude)	1	1.7266	0.08	-
(Change in Knowledge)	1	12.8465	0.58	-
Participation	1	200.9286	9.06**	.0799
Training	1	3.8709	0.17	-
Part x Train	1	0.2120	0.01	-
Residual	90	1996.7091	-	-
Total	95	2216.2937	-	-

Table 26. Study Ten MANOCOV and ANOCOV Tables (Cont'd)

Note. The dependent variables in these analyses are the variances of the subjects' ratings, across professors, for each category.

[a]The F tests for the MANOCOVs were approximated by the Hotelling-Lawley and Pillai traces.

*p < .05

**p < .01

or in combination with Change in Knowledge did not "wash out" the significant multivariate Participation x Training interaction effect, nor did it affect the significance of the interaction test for any of the univariate tests. The only effect of covarying Change in Attitude, either separately or in combination with Change in Knowledge, was to allow the Participation main effect to reach statistical significance. Presumably the trend for Participation to lead to an undesirable increase in dispersion scores for Category E, noted in Study Three, became strong enough to influence the multivariate test. As in Study Eight, none of the covariance effects were found to be statistically significant.

Discussion

The results of this study, taken in conjunction with those of Study Eight, clearly do not support the predictions of the hypothetical model. Variables other than Knowledge and Attitude apparently mediate the effects of Participation and Training on various characteristics of ratings. Chapter V considers in detail the implications of these findings for the hypothetical model, and includes an attempt to identify variables which may be possible mediators.

CHAPTER V

DISCUSSION AND IMPLICATIONS

The major conclusion of this investigation from the applied perspective appears to be that gains in psychometric characteristics of performance ratings can result from allowing raters to participate in constructing the rating scales they are to use. The finding that Participation can significantly increase the reliability and validity of ratings--especially in view of the fact that the reliability and validity scores reported in this study compare quite favorably to what Borman (1978) has claimed are the upper limits of reliability and validity in job performance ratings--lends strong support to the recommendation voiced by Smith and Kendall (1963) and others (e.g., Bernardin, Alvares, & Cranny, 1976; Borman & Vallon, 1974; Campbell et al., 1973; Friedman & Cornelius, 1976) that participative techniques be considered by any practitioner who is seeking to construct adequate performance rating scales. The present investigation provides empirical support for these recommendations.

This investigation generated little support for the choice of Training over Participation as a technique for improving the psychometric quality of ratings. Participation outperformed Training in improving every characteristic of the ratings except overall elevation (leniency). If, however, leniency error is perceived as the major problem affecting quality of performance ratings in the organization in

question, then the practitioner may be justified in choosing a simple training program over a more complex, time consuming participative effort--especially in view of Borman's (1975) finding that even a five-minute training program can be somewhat effective in terms of error reduction.

This investigation did not consider the two treatments from a cost-benefit perspective, yet these concerns must be weighed by the practitioner. The participation program undertaken in this investigation required almost twice as much time from the subjects as did the training program, and the cost in terms of the experimenter's and his supportive colleagues' time was far greater for the participation than for the training program. The practitioner must consider the relative costs of two types of programs in comparison to the possible financial benefits of an improved performance appraisal rating system--or the possibly substantial costs of losing a civil rights-related case in court due to a poorly designed criterion measurement system. Perhaps, as suggested by Brogden and Taylor (1950a) and Mirvis and Lawler (1977), a broader approach to measuring the financial impact of a poor performance evaluation system should be considered, especially in light of Glickman's (1955) and Bass' (1956) demonstrations of the detrimental effects of poor criterion measurement techniques on employee motivation and personnel administration.

Several of the questions posed at the end of Chapter I dealt with the interaction of Participation and Training. The findings of most of the studies reported in Chapter IV indicate no significant interaction

effects among Participation and Training, suggesting that these two treatments independently influence characteristics of ratings. However, it was found that Participation and Training combined interactively to influence Knowledge scores and the dispersion of ratings per category. Perhaps, viewing the findings of this investigation in the context of the results of the studies discussed in Chapter I, the most effective treatment for reducing error and enhancing the reliability and validity of ratings would be a combination of the participation and training approaches. A program consisting of rater participation in scale construction, "sandwiched" between two training sessions--the first dealing with philosophical issues in performance rating, the second with error reduction techniques and practice in actual use of the scales--appears worthy of investigation as an approach to establishing an effective performance rating system.

In addition to their important practical implications, the findings of this investigation also have general importance for psychological measurement. The present findings, supported by those of earlier studies, demonstrate that involving experimental subjects in the development of the measurement scales they are to use may lead to a great improvement in the quality of the subjects' evaluations of characteristics of conceptually defined stimuli. This finding is certainly important to industrial-organizational psychologists, who are frequently required to measure difficult-to-define constructs in a milieu of uncontrolled sources of variance. However, industrial-organizational specialists are not the only psychologists dealing with problems of measurement in field

studies. This situation is common to many specialties of psychology, including clinical, educational, developmental, social, and experimental psychology. The extent to which evaluations of behavior and of various constructs in such areas, gathered via the method of single stimuli, might be made more psychometrically sound by involving in the construction of the measurement instruments persons who are to make the evaluations seems deserving of experimental investigation. By becoming directly involved in the attempt to measure an ambiguous stimulus, participant subjects may increase their understanding of the stimulus, and thus reduce some of the ambiguity, and constant error, in its measurement. A research project investigating the effects of participation in scale construction on the quality of measurements taken in contexts ranging from the hospital psychiatric ward and the elementary school classroom or playground to the industrial or consumer setting, or even to the psychophysics laboratory, is envisioned. Perhaps investigations of the effects of Participation and Training in these varied contexts will yield information upon which to develop a clearer understanding of the effects of these treatments on the quality of the measurements obtained.

Implications for the Hypothetical Model

Simply stated, the hypothetical model proposed in Chapter II suggests that Participation and Training (treatments) alter raters' attitudes toward and knowledge about the performance appraisal rating process (mediators), and that Attitude and Knowledge, in turn, affect psychometric characteristics of performance appraisal ratings (outcomes). The model

further suggests that Attitude and Knowledge are the sole mediators of the effects of the treatments on the outcomes. The three phases of the data analysis employed in this study were intended to examine these three major propositions of the hypothetical model. It is now appropriate to evaluate and reformulate the hypothetical model on the basis of the findings of this investigation.

The first major proposition of the model is that the two treatments, Participation and Training, influence various psychometric characteristics of performance appraisal ratings. This hypothesis is strongly supported for the Participation treatment, as predicted by such authors as Borman and Vallon (1974), Campbell et al. (1973), and Friedman and Cornelius (1976). In the present investigation, Participation was found to significantly increase estimates of discriminant validity, intraclass and one-rater reliability, and intraclass and one-rater validity of the set of ratings, while significantly reducing an estimate of consensual halo error.

Counter to the predictions of Bernardin and Walter (1977), Borman (1975), Dickinson and Tice (1973), Guilford (1954, p. 280), and Zedeck and Baker (1972), major training effects were not found in this investigation. While Training did produce a statistically significant decrease in leniency error, it accounted for only a small proportion of variance in the ratings. Furthermore, Training was not found to affect estimates of the reliability or validity of the ratings. While skeptics might argue that the rater training program possibly was not properly designed or presented, an examination of the content of the program

(see Appendix H) reveals its strong similarity to other programs reported effective in influencing characteristics of performance appraisal ratings (e.g., those of Bernardin & Walter, 1977; Bittner, 1948; Brown, 1968; Latham et al., 1975; and Stockford & Bissell, 1949). The finding that Training significantly increased Knowledge also disputes any such criticism of the training program.

In view of the many studies reporting successful application of rater training programs (see Chapter I), it is not appropriate to eliminate, on the basis of the results of this single study, rater training as an option for increasing the quality of performance appraisal ratings. It is, however, appropriate to conclude (for at least the conditions in effect in this study) that participation in scale construction appears to be a key to improving the psychometric quality of performance appraisal ratings. Perhaps with subjects who are less knowledgeable about the performance appraisal rating process, or who have less favorable initial attitudes toward rating, Training would be far more effective. The likelihood of finding stronger Training effects might also increase if scales of the type more commonly used in the field were employed as the appraisal instrument, rather than the carefully constructed BARS used herein. These speculations are, of course, appropriate for future investigation.

The second major proposition of the model is that the treatments, Participation and Training, affect measures of attitude toward and knowledge of the performance appraisal rating process. This proposition is supported by the data reported in this dissertation. Participation did

affect change in Attitude, although not in the direction anticipated. Numerous authors (e.g., Barrett, 1966, p. 14; Bittner, 1948, p. 423; Coch & French, 1948; Friedman & Cornelius, 1976; Levine & Butler, 1952; Rundquist & Bittner, 1950; Smith & Kendall, 1963) suggest that Participation will increase favorability of attitudes toward performance appraisal, yet this investigation found the opposite effect. As noted in Chapter IV, the most likely explanation of this finding is in terms of Lowin's (1968, p. 80) hypothesis that participation followed by no visible change in organizational policy or procedure may not be at all motivating. In fact, Strauss (1963, p. 70; Bavelas & Strauss, 1961, p. 590) believes that employee attitudes will be strongly adversely affected by this condition. In this experiment, subjects were informed that the scales were being developed for experimental purposes only and would not be used administratively; thus the conditions were appropriate for a possible "backfire" effect of Participation on attitudes. This possible "backfire" effect of participation in decision-making followed by no visible attempt to implement the consensus decision certainly deserves additional investigation.

The model predicted that the rater training program would significantly increase raters' attitudes toward the performance appraisal rating process. The data did not support this hypothesis--Training did not significantly affect Attitudes. Perhaps the base attitude level of the student subjects, which was in the "slightly favorable" range, was already too high to be significantly influenced by Training. Research

with groups of subjects at various levels of initial attitude toward the rating process might clarify this issue.

Training did, however, clearly increase Knowledge, as predicted on the basis of the findings of studies summarized in Chapter I. The statistically and practically significant effect of Training on Knowledge appeared despite the relatively high level of pre-treatment Knowledge exhibited by the subject sample. It is quite likely that the effect would be much stronger with a subject sample whose initial level of Knowledge was not as restricted in range. The model did not predict a significant effect of Participation on Knowledge, and none was found. It is interesting, however, that Training was more effective in increasing Knowledge scores for Participant than for Non-participant subjects. This deserves more attention. As noted in Chapter IV, it is possible that this is a manifestation of the increased motivational level expected to result from participation in scale construction.

The third major prediction of the hypothetical model examined in this investigation dealt with the mediating effects of Attitude and Knowledge. The model suggests that Attitude and Knowledge are the sole mediating variables of the effects of the treatments on characteristics of the ratings. Thus, if the model were correctly formulated, any effects of the treatments on such characteristics could be expected to disappear when the treatments' effects on the two mediators are held constant through covariance analysis. The findings of Studies Eight and Ten refute this portion of the model. If Attitude and Knowledge mediate the effects of the treatments at all, they are certainly not

the sole mediators. Speculation regarding additional mediating variables appears below. Before considering these variables, however, it is appropriate to present a revised hypothetical model of the effects of Participation and Training on characteristics of performance appraisal ratings (see Figure 4).

In the revised model, solid lines represent strong effects; dashed lines represent weak effects. The revised model suggests that Training and the Participation x Training interaction affect Knowledge. The new model also reflects the belief that Participation and Training influence psychometric characteristics of ratings. The new model retains the links between Attitude and Knowledge and characteristics of ratings, although little support for these links has thus far been found. The major change from the original hypothetical model is the provision for additional mediating variables (discussed below). The model is, of course, incomplete--it should be viewed as a subsystem of a larger model rather than as a closed-system entity. As demonstrated in Table 1, many other variables influence characteristics of ratings. It is quite likely that other variables influence the mediators as well.

An expanded, more complete model would include characteristics of the raters, such as initial levels of Attitude and Knowledge. Several contextual or situational variables would also be included in a more complete model. Some of the major candidates for inclusion are management's expressed concern for the rating scheme, visibility of efforts to implement changes suggested by Participant subjects, subjects'

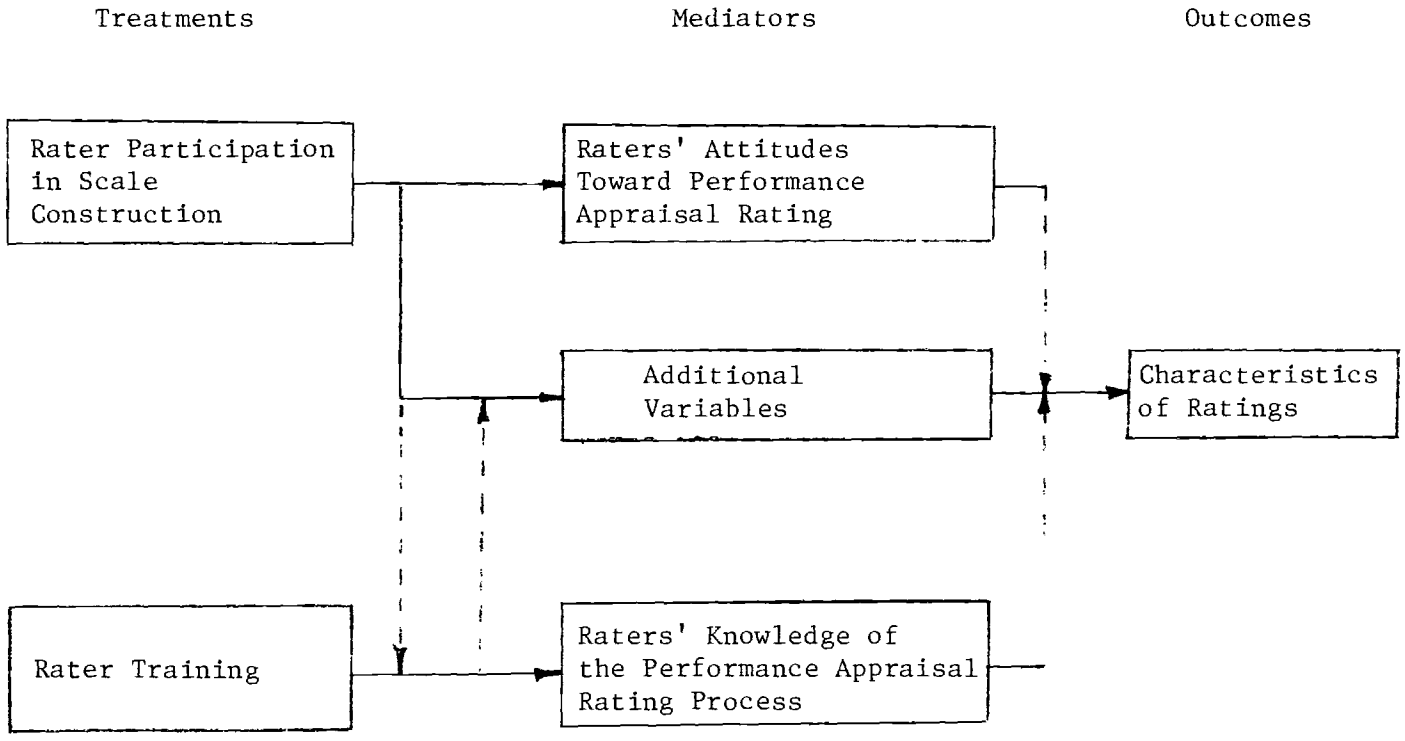


Figure 4. A Revised Hypothetical Model of the Effects of Training and Participation

perceptions of whether outcomes of their instrument-construction efforts will be implemented, and raters' perceptions of the results of their appraisals of subordinates. An expanded model might also consider outcomes of Training and Participation other than effects on psychometric characteristics of ratings. Possible "side effects" might include attitudes toward being rated as well as toward making ratings, perceptions of organizational demands and expectations, and perceptions of the importance of and management's concerns for the performance appraisal process.

The challenge with regard to the new hypothetical model presented in Figure 4 is the identification of the additional mediating variables. One possible candidate has already been suggested: attitude or motivation toward the use of the specific rating scale or instrument in question. An individual may be strongly in favor of performance appraisal in the abstract, but may be quite opposed to the use of a particular rating instrument. Conversely, the individual may not be strongly impressed with the idea of performance rating, but may be motivationally committed to the proper and effective use of a certain rating scale. Such a motivational commitment has often been cited as one outcome of participative decision making (Lowin, 1968; Maier, 1967; Vroom & Yetton, 1973; Wood, 1973). Perhaps, then, both general attitude toward performance appraisal rating and specific attitude toward the particular rating instrument in question should be assessed and examined as separate mediating variables.

Schneier (1977) has recently suggested a second possible mediator of the treatment-outcome relationship. He presented evidence indicating that "cognitive complexity," defined as "the degree to which a person possesses the ability to perceive behavior in a multidimensional manner" (p. 541), affects subjects' ratings of confidence in and preference for BARS as opposed to rating scales with a simpler format. Schneier also reports that cognitive complexity affects leniency and central tendency errors in BARS ratings. Finally, he suggests that "future research employing a longitudinal design could ascertain whether participation in [BARS] development and use increases cognitive complexity" (p. 547). A study recently completed by Pond and Sauser (Note 11) employing the longitudinal methodology suggested by Schneier failed to support the hypothesis that either Participation or Training affects cognitive complexity scores. Nor was initial level of cognitive complexity found to moderate the effects of Participation or Training on Attitude change, Knowledge change, or extent of leniency and central tendency errors in BARS ratings. Bernardin and Boetcher (Note 9) also report a failure to substantiate Schneier's claims of cognitive complexity as a moderator of the effectiveness of BARS use.

A third potential mediating variable is suggested in the writings of Barrett (1966, p. 121), Bittner (1948, pp. 422-424), Brown (1968), Dunnette (1966, p. 100), Friedman and Cornelius (1976), King et al. (1952), Thorndike and Hagen (1969, p. 446), and Smith and Kendall (1963). These authors speculate that one effect of Training and/or Participation is to help clarify for the raters the behavioral categories and anchor

points included in the rating instrument. They suggest that the clearer the rater's understanding of, and ability to differentiate among, categories and anchor points on the rating scales, the better are the ratings produced. Note that one point of emphasis in the training program included in this investigation (see Appendix G, Part II), as suggested by Bittner (1948) and Brown (1968), was instruction on the meaning of the categories to be rated and the anchor points used in the BARS. This instrument-specific material stands apart from the content of the training program dealing with general knowledge of the performance appraisal rating process, and is not covered in the Knowledge Scale. Clarity of understanding of scale categories and anchor points appears to be a potential mediating variable worthy of additional investigation.

Unfortunately, this "clarity" variable appears to be quite difficult to measure. An instrument recently designed for this purpose by the author proved to have adequate content validity, yet reliability estimates for the scale ranged from .25 to .62. Bernardin (Note 12) reports similar difficulty in his attempts to operationalize this construct. Evidence presented by Zedeck and Baker (1972; see also Zedeck et al., 1976) suggests that unreliability in measures of the construct may be due at least in part to true instability of the skill necessary to make differentiations among behavioral categories and anchor points. Nonetheless, it is quite possible that this skill may prove to be an important mediator of the treatment-outcome relationship specified in the hypothetical model.

Limitations of the Investigation

The major limitations to the generalizability of the findings of this investigation stem from the experimental control-generalizability tradeoff. This project was intended to examine relationships among a small set of variables under closely controlled conditions. There are a plethora of variables which may affect the outcomes of a rating scale evaluation study. To complicate matters, many of these variables may combine interactively to moderate the effects of the two treatments examined in this study. Given the strict controls over the nature of the experimental subjects, the rating context and stimulus materials, and the characteristics of the rating instrument employed in this investigation, it is improper to attempt to generalize across all possible levels of these variables. Nonetheless, the findings of this investigation yield suggestions for scale development and use which can be tested in various other contexts. The finding, for instance, that Participation appears to be an agent affecting psychometric characteristics of ratings gathered in a controlled situation lends credence to claims for similar results found in field research settings where many of these controls are relaxed.

Several specific limiting characteristics of this investigation have been mentioned earlier. Certainly one such factor is the simulated nature of the stimuli to be rated. The diaries eliminated a major source of variation in ratings by ensuring that all subjects attended to identical sets of behaviors. Factors which might strongly affect the raters' decision-making processes, such as the amount of information to which

they are exposed (Einhorn, 1971), or their freedom to choose which behaviors to observe, were not allowed to vary. This is certainly not the case in most applied settings, as pointed out in Chapter I. Perhaps an operational replication of this study under conditions where standardized stimuli are observed in situ via film or videotape would help bridge the generalization gap between this investigation and the typical applied setting. A history of informal, day-to-day interactions between raters and ratees characterizes many applied settings. It may be speculated that this history of interactions leads to much of the psychometric error found in ratings gathered in the field. Although the diary approach employed in this experiment provided a flavor of this history of interactions, the diaries cannot be claimed as adequate substitutes for the many hours of raters' exposure to ratees which is found in many applied settings. Unfortunately, this exposure may not be reproducible other than with elaborate, lengthy simulations. "Opportunity to observe" should, of course, continue to be investigated as a moderator of the effects of treatments on characteristics of ratings (Ferguson, 1949a; Ghiselli & Brown, 1955, p. 90; Smith, 1976, p. 762; Thorndike & Hagen, 1969, pp. 427-428).

The apparent high psychometric quality of the BARS employed in this investigation has already been mentioned as a possible factor limiting the generalizability of the results of this experiment. As Kipnis (1960), Ronan and Schwartz (1974), Taft (1955), and Toops (1944) have pointed out, very few of the rating schemes currently used in industry have been developed as carefully as those employed in this

experiment. Perhaps had more typical scales been used, the positive effects of Training would have been apparent. Another factor which might limit the generalizability of this investigation's findings regarding Training is the initial positive attitude toward rating scales, and high level of sophistication with respect to their use, exhibited by the experimental subjects. Certainly systematic replications of this investigation with subjects other than college students would help establish the population validity (Anderson, Ball, Murphy, & Associates, 1975, p. 459) of these findings.

It must be recognized that professors represent "significant others" to college students, since professors have a great deal of control over the behavior of students through the mechanisms of reward and sanction. In the typical applied setting, the rater would most likely be in the position of authority, not the ratee as in this experiment. The possible effects of this reversal in the typical rater-ratee relationship should be investigated as a potential limiting factor. Again, systematic replications of this experiment with additional subject samples would be beneficial.

The ecological validity of the findings (Anderson et al., 1975, p. 459) also deserves examination through systematic investigation across contexts and environments. The applied versus research context is one potential moderator of these findings which deserves further analysis (Taylor & Hastman, 1956; Taylor & Wherry, 1951).

Suggestions for Future Research

Throughout this report the author has made suggestions for future research which may help clarify the myriad of issues surrounding the effective use of performance appraisal rating instruments. The purpose of this final section is to collect some of these thoughts in an attempt to map directions for research. One suggestion is to systematically replicate this investigation across various subject samples and environmental contexts, perhaps varying several of the moderating varibles mentioned in the previous section, in order to get a clearer idea of the true effects of Training and Participation on characteristics of ratings. An extension of the research from educational to other settings and contexts of interest to psychologists is also recommended.

A second suggestion involves systematically investigating the nature of the two treatments dealt with in this project. Bittner (1948) and Brown (1968) have attempted to delineate the essential content of a rater training program. Similarly, Bernardin, LaShells, Smith, and Alvares (1976), Dickinson and Tice (1973), and Friedman and Cornelius (1976) have grappled with the issue of identifying the required components for a successful rater participation experience. Both of these areas deserve additional investigation. Perhaps an examination of the extent of participation (direct or representative; see Coch & French, 1948; Lowin, 1968) necessary to produce psychometrically sound rating instruments would yield information valuable from both the theoretical and practical points of view. It would also be of interest to further explore the relationship between participation and attitude change,

especially in the situation where the outcomes of the participative sessions were not implemented (Bavelas & Strauss, 1961, p. 590; Lowin, 1968, p. 80; Strauss, 1963, p. 70). A study equating the time spent under each treatment might also be beneficial in exploring the effectiveness of participation and training.

This investigation suggests that change in general attitudes toward or knowledge about the rating process does not account for all of the effects of the treatments on characteristics of ratings. According to Pond and Sauser (Note 12), neither does change in cognitive complexity, as hypothesized by Schneier (1977). As indicated above, more research on possible mediators, especially clarity of understanding of scale categories and anchor points and specific attitudes toward scale usage, needs to be undertaken. Cost/benefit analyses of Participation, Training, and the other methods intended to improve the pyschometric quality of ratings are also called for. It might also prove beneficial to investigate possible "side effects" of Participation and Training, such as changes in attitudes toward being rated and changes in perceptions toward organizational demands and expectations, especially with regard to the performance appraisal process.

Numerous interesting studies might be performed with the BARS and other materials constructed for use in this project. Following the lead of Cascio and Valenzi (1978), Dickinson and Tice (1973), and Zedeck and Baker (1972), it may be valuable to examine relationships among BARS and "hard" criteria of professional performance. A comparison of the incident evaluations of students, professors, and administrators,

as suggested by Tauscher's (Note 3) study, may also be useful, especially in view of the policy implications suggested by Blood (1974). By systematically varying characteristics of the simulated professors, it may be possible to study the biasing effects on student ratings of such variables as sex, race, age, rank, and reputation of the ratee under tightly controlled conditions. Rater x ratee interactions (see Table 1, section IX) might also be systematically examined in this manner.

Another interesting avenue of research is suggested by the significant Category and Participation x Category effects found in this investigation. The hypothesis that ratings of less well defined, less observable categories of behavior might be more susceptible to distortion through the halo and logical errors (Symonds, 1925) certainly deserves attention. The possible moderating effects of Category on the effectiveness of the other methods of reducing psychometric error in ratings mentioned in Chapter I also deserves further investigation.

Many other topics for research are suggested by the taxonomy of sources of variance and error in ratings (Table 1) and model of the rating process (Equation 2) presented in Chapter I. As Friedman and Cornelius (1976) point out, a systematic analysis of the various factors that influence ratings, especially the interactive factors, is "sorely needed" (p. 216). This investigation is but a small step in that direction.

APPENDICES

A. Informed Consent Forms

A-1. Program Evaluation Study
A-2. Questionnaire Development Studies

B. Attitude Scales

B-1. Attitude Scale Form A
B-2. Attitude Scale Form B
B-3. Item Statistics for the Attitude Scales

C. Knowledge Scale

D. Criterion Scale

E. Critical Incident Reporting Forms

E-1. Category of Teacher Behavior
E-2. Positive Incident in Adolescence
E-3. Negative Incident in Adolescence

F. Behaviorally Anchored Rating Scales

F-1. BARS for Categories A, B, C, D, E, with Instructions
F-2. BARS Item Statistics

G. Simulated Professors

G-1. Professors L, M, N, O, P
G-2. Item Statistics for Simulated Professors

H. Rater Training Program

APPENDIX A-1

Program Evaluation Study

Informed Consent Form

I understand that:

(1) The general purpose of this study is to evaluate several alternative psychological programs. It will require me to attend five two-hour sessions held on five consecutive Mondays or Tuesdays. Failure to attend any session will constitute withdrawal from the study.

(2) I will be asked to fill out several questionnaires in the first session, to describe and evaluate numerous examples of behavior and/or to receive classroom training during the middle three sessions, and to fill out several more questionnaires in the last session.

(3) My questionnaire responses will be kept strictly confidential. No individuals will be identified if the results of this study are published or otherwise disseminated.

(4) Some of the behavior descriptions I turn in anonymously may be reviewed by other students. However, I have the right to request that they not be seen by others if I so desire.

(5) I am free to withdraw from this study at any time.

(6) I will be "debriefed" at the conclusion of the fifth session. Also, I am entitled to a full explanation of the results of this study when they are available, and I may obtain this information from Professor Sauser (HC 1224-B) if I so desire.

(7) I will receive 10 hours of experimental credit for participating in this study in addition to a chance at a cash prize of $50, one prize per group of 25-30 subjects to be awarded by lottery at the conclusion of the fifth session.

(8) The outcomes of this scientific study are based on the assumption that I will honestly and conscientiously perform the tasks required of me.

APPENDIX A-1 (Cont'd)

(9) The nature of this study requires that I work independently and that I will not discuss any part of the study with anyone until the fifth session has been completed and the experimenters tell me that I may.

I have read this form and voluntarily consent to participate in the Program Evaluation Study.

_____
Subject's signature

_____
Professor's name

_____
PG class and meeting time

APPENDIX A-2

Questionnaire Development Study

Informed Consent Form

I have had the purpose and general nature of this study explained to me and I consent to participate in it. I realize that my responses will be kept strictly confidential and that no individuals will be identified if and when the results of this study may be published or otherwise disseminated. I understand that I am free to withdraw from this study at any time. I also realize that I am entitled to a full explanation of the results of this study when it is completed and that I may obtain this information by contacting Prof. Sauser if I so desire. I understand that I will receive one hour[1] of experimental credit for participating in this study. I realize that the outcomes of this scientific study are based on the assumption that I will complete these forms honestly and conscientiously. I am also aware that the nature of the study requires that I not discuss my responses to these questionnaires with anyone until the experimenter tells me that I may.

_____

Subject's signature

---

[1]The amount of credit awarded differed from study to study.

APPENDIX B-1

Scale 2, Form A

Directions: Read each of the statements below and decide whether or not you agree with it. If you agree with the statement, circle A next to the item number. If you disagree with it, circle D. If you are undecided, circle U. Be sure to circle either A, U, or D for each of the 30 statements.

A U D 1. I hope I never see another teacher rating form.

A U D 2. Teacher evaluations often change professors' behavior for the worse.

A U D 3. I think faculty evaluation is extremely valuable and important.

A U D 4. Teacher rating systems should be avoided at all costs.

A U D 5. Teacher evaluation sometimes seems like a good idea.

A U D 6. Teacher rating forms are too hard to fill out.

A U D 7. I don't usually mind filling out teacher rating forms.

A U D 8. I can see how faculty evaluation might do some good.

A U D 9. A good teacher rating system can help solve many of a college's problems.

A U D 10. A teacher rating form is an extremely valuable tool for evaluating performance.

A U D 11. Questions on some teacher rating forms are hard to interpret.

A U D 12. Most students don't mind rating their professors.

A U D 13. Some questions on faculty evaluation forms seem irrelevant.

A U D 14. Faculty evaluation is a great idea.

A U D 15. Faculty evaluation forms provide a useful standard for teacher performance.

A U D 16. Faculty evaluation is sometimes worthwhile.

APPENDIX B-1 (Cont'd)

A U D 17. Most teacher evaluation forms are thorough and relevant.

A U D 18. Performance rating scales are convenient forms for evaluating teachers.

A U D 19. It is sometimes inconvenient to evaluate professors.

A U D 20. I am very much opposed to faculty evaluation.

A U D 21. Teacher ratings are more trouble than they are worth.

A U D 22. Teacher evaluation is tremendously important.

A U D 23. Teacher performance rating is simply a waste of time.

A U D 24. I hate teacher rating forms.

A U D 25. Teacher rating systems do more damage than good.

A U D 26. I can't think of anything more important to a university than faculty evaluation.

A U D 27. Faculty evaluation doesn't seem important to most students.

A U D 28. Teachers are opposed to faculty evaluation.

A U D 29. Most professors are glad to get information from their students.

A U D 30. Teacher evaluation doesn't make a whole of sense to me.

APPENDIX B-2

Scale 2, Form B

Directions: Read each of the statements below and decide whether or not you agree with it. If you agree with the statement, circle A next to the item number. If you disagree with it, circle D. If you are undecided, circle U. Be sure to circle either A, U, or D for each of the 30 statements.

A U D 1. Teacher rating forms are not worth the paper they are printed on.

A U D 2. Teacher evaluation forms are more valuable than gold.

A U D 3. Rating forms ensure that all teachers are evaluated fairly.

A U D 4. Teacher evaluation forms provide useful feedback to professors.

A U D 5. Some teacher rating forms can be improved.

A U D 6. Faculty evaluation doesn't seem to do much good.

A U D 7. Teacher evaluation results are sometimes useful.

A U D 8. Faculty evaluation is a useless waste of time.

A U D 9. Teachers sometimes learn things from their evaluations.

A U D 10. Teacher evaluations serve little purpose in today's academic world.

A U D 11. There are no real benefits of teacher evaluation.

A U D 12. Rating their professors is the most important thing students do in college.

A U D 13. Teacher evaluation questionnaires often seem redundant.

A U D 14. Teacher evaluation forms are sometimes too long.

A U D 15. Faculty evaluations can hurt professors' feelings.

A U D 16. Teacher rating forms are absolutely fantastic.

APPENDIX B-2 (Cont'd)

A U D 17. The process of teacher evaluation builds a more qualified faculty.

A U D 18. Colleges should not be allowed to use teacher rating forms.

A U D 19. Faculty evaluation is not really necessary.

A U D 20. Faculty evaluation forms are extremely useful.

A U D 21. I don't mind spending a little time rating my teachers.

A U D 22. Teacher ratings sometimes do some good.

A U D 23. Teacher evaluation forms are worthless.

A U D 24. Teacher evaluation is essential if a college is to maintain high standards.

A U D 25. Teacher evaluation leads to more effective instruction.

A U D 26. Too much time is spent on faculty evaluation.

A U D 27. Teacher rating forms sometimes ask for too many details.

A U D 28. Teacher rating forms are generally unpopular.

A U D 29. Most faculty rating forms don't ask for enough information.

A U D 30. I sometimes think teacher rating forms are useful.

APPENDIX B-3

ITEM STATISTICS FOR THE ATTITUDE SCALES

Form A			Form B		
Item #	S	Q	Item #	S	Q
24	1.3	0.9	8	1.3	1.1
1	1.3	1.2	23	1.3	1.2
20	1.4	1.1	1	1.4	1.3
4	1.5	1.3	18	1.5	1.0
23	1.9	1.8	11	2.3	1.7
21	2.4	1.6	10	2.4	1.4
25	2.6	1.5	6	2.6	1.5
2	3.0	1.5	19	3.0	1.5
6	3.2	1.2	26	3.4	1.4
30	3.6	1.8	28	3.6	1.7
28	3.6	1.9	13	4.2	1.6
27	4.3	1.4	29	4.2	1.8
13	4.3	1.5	14	4.3	1.8
11	4.6	1.9	27	4.5	1.6
19	4.7	1.5	15	4.8	1.9
7	6.4	1.6	5	5.4	1.7
12	7.2	1.7	30	7.3	1.1
5	7.3	1.5	21	7.3	1.3
16	7.4	1.3	22	7.4	1.6
8	7.6	1.5	7	7.5	1.5
18	7.6	1.7	9	8.3	1.4
17	8.5	1.7	4	8.3	1.6
29	8.6	1.6	3	8.6	1.9
15	8.7	1.8	17	8.6	1.9
9	8.8	1.9	25	9.0	1.7
10	9.4	1.7	20	9.2	1.5
14	9.6	1.7	24	9.7	1.8
22	10.2	1.2	12	10.0	1.8
3	10.2	1.3	16	10.3	1.5
26	10.6	1.3	2	10.5	1.3
Mean	5.73	1.52	Mean	5.74	1.54
S.D.	3.07	0.25	S.D.	3.03	0.25

APPENDIX C

Scale 1

Directions. This questionnaire consists of 150 true-false items. Please read each item and circle either T (if you think the statement is true), or F (if you think it is false).

T F 1. In order to enhance learning, professors should wait at least one week before returning corrected examinations.

T F 2. A student in residence at Auburn University may not enroll in a correspondence course if the course or a suitable substitute can be scheduled.

T F 3. The University Placement Service charges a $25 non-refundable fee to any student seeking its assistance in finding post-graduation employment.

T F (4.) Faculty evaluation forms are a systematic way to gather information about teacher performance.

T F (5.) There is uniform agreement on the interpretation of "good teaching performance."

T F (6.) Teacher evaluation results can be used to influence decisions regarding course assignments and class sizes.

T F 7. Course credits earned by special students generally cannot be used toward a degree at Auburn University.

T F 8. In a learning situation, the learner should be helped to evaluate his own performance.

T F (9.) The most important thing to remember when filling out teacher evaluation forms is to be consistent from item to item.

T F 10. An Auburn University student may request that any information contained in his educational records which he considers to be inaccurate or misleading be amended or deleted from the records.

T F (11.) There are many different duties involved in the job of college professor.

T F 12. A normal load for Auburn University undergraduate students is 10-14 hours per quarter.

T F 13. Children typically learn better when they practice tasks as a whole.

T F (14.) When rating a professor, it is best to form a general, overall impression of his performance, then let your responses reflect this impression.

T F 15. Auburn University is over 100 years old.

T F 16. Auburn University requires a minimum of 20 quarter-hours in Natural Science for the bachelors degree.

Note. The 50 items making up the Knowledge Scale have been circled and the answers keyed. The other 100 items are the "Irrelevant" Knowledge Scale.

APPENDIX C (Cont'd)

T F (17.) Teacher evaluation is a relatively effective way to provide feedback to professors.

T F 18. Individuals may apply to Auburn University for entrance to any quarter of a calendar year as early as October 1 of the preceding year.

T F 19. Auburn University does not provide a formal pre-college counseling program.

T F 20. Auburn University provides over 40 residence halls for its students.

T F (21.) Each item on a faculty rating scale should be considered independently.

T F (22.) Teacher evaluation provides a way for professors to become aware of their strengths and weaknesses.

T F 23. Auburn University will release a student's educational records upon the student's written request.

T F 24. Children should be encouraged to practice the skills they have learned.

T F (25.) It is very difficult for any one measure to take into account all the facets of teaching performance.

T F (26.) You should try not to let your friendship with a professor affect your rating of him.

T F (27.) There are a number of ways that faculty evaluation information can be used.

T F (28.) Informal means of faculty evaluation are often biased.

T F 29. Applicants of mature age who are not high school graduates may be considered for admission at Auburn University if their educational attainments are shown through testing to be equivalent to those of a high school graduate.

T F 30. A student with a baccalaureate degree who undertakes a program for a second bachelor's degree at Auburn University is classified as a graduate student.

T F 31. The Auburn University Computer Center is located on the first floor of Parker Hall.

T F 32. Grades earned in correspondence courses are included in the Auburn University grade point average.

T F 33. Students who have served in the Armed Forces may receive credit at Auburn University for military courses completed at the college level.

APPENDIX C (Cont'd)

T F 34. Women were first admitted to Auburn University sixty years ago.

T F 35. Teacher evaluations must be consistent and reliable to be useful.

T F 36. A minimum of sixty hours must be earned in residence at Auburn University in order to receive a bachelor's degree.

T F 37. Professors often informally evaluate each others' teaching performance.

T F 38. Auburn University offers courses comparable to high school geometry and first and second year high school algebra.

T F 39. Unnecessary errors should be avoided in a learning situation.

T F 40. Professors teaching difficult courses should not be evaluated any differently than those teaching relatively easy courses.

T F 41. Deans and department heads usually observe and evaluate their professors' teaching performance first-hand.

T F 42. Each student becomes a member of the Student Government Association upon enrollment at Auburn University.

T F 43. Professors are evalauted informally as well as formally.

T F 44. Auburn University requires a minimum of nine quarter-hours in English composition for the bachelors degree.

T F 45. Age and sex should not affect students' ratings of professors.

T F 46. What one student believes is "excellent" teaching performance may be seen as only "average" by another.

T F 47. Auburn University gives preference in undergraduate admissions to residents of Alabama.

T F 48. Auburn University was closed during the Civil War.

T F 49. The Auburn University Chapel is open only on weekends.

T F 50. "Cramming" is typically an effective way to learn material thoroughly.

T F 51. Dislike of school does not typically affect children's efforts to learn.

T F 52. You should check with your classmates and see what they think before filling out a rating form on your professor's teaching performance.

APPENDIX C (Cont'd)

T F 53. Auburn University has approximately 60 major buildings on its campus.

T F 54. At Auburn University, no penalty shall be assigned for dropping a course during the first four weeks of the quarter.

T F 55. Errors which contribute knowledge of results sometimes lead to improved learning.

T F (56.) When rating professors, it is best to rate them somewhere in the middle of the scale, avoiding the extremes.

T F 57. The largest single source of Auburn University revenues is state appropriations.

T F (58.) Most faculty evaluation forms are constructed by concerned students.

T F 59. A student of high academic promise may be admitted to Auburn University directly from the eleventh grade without a diploma.

T F (60.) The type of form used can affect the outcome of a teacher evaluation project.

T F 61. Auburn University's academic program is fully accredited by the Southern Association of Colleges and Schools.

T F 62. The Auburn University Student Health Service is under the direct supervision of the Dean of Student Services.

T F 63. Rest breaks should be avoided when studying for examinations.

T F 64. Auburn University students have the right to inspect and review the contents of records directly relevant to the student.

T F (65.) A professor's appearance should not typically influence the way students rate his teaching performance.

T F 66. At Auburn University, a grade of 0(zero) or F may be assigned for academic dishonesty.

T F 67. At Auburn University, students are not permitted to change from S-U to conventional grading after the schedule adjustment period.

T F 68. Auburn University's Cooperative Extension Service provides service to 133 counties in Alabama.

APPENDIX C (Cont'd)

T F (69.) When rating a professor's teaching ability you should keep in mind the quality of his research.

T F 70. Auburn University may refuse admission to any individual whose health record indicates that his health or the University community might be adversely affected by his attendance.

T F (71.) It is important that faculty evaluation items be designed such that each student interprets them a little differently.

T F (72.) When rating a professor whose class you have not attended regularly, you should base your ratings on what your friends have said about the professor.

T F 73. In a learning situation, errors should typically be corrected upon their first appearance.

T F 74. A student may earn a maximum of 25 percent of the total credits required for his Auburn University baccalaureate degree by correspondence or extension.

T F 75. All non-Alabama resident Auburn University students except graduate students are required to pay a tuition fee.

T F 76. Knowledge of results inhibits learning.

T F 77. Clearly informing the learner of what to do is essential if learning is to take place.

T F 78. The Dean of Student Life serves as the social director of Auburn University.

T F (79.) When rating professors it is best to use the extreme ends of the scale on every item.

T F (80.) The professor's reputation in the field should not be considered when the student is filling out a faculty evaluation form.

T F 81. Every student who makes use of the instructional staff and facilities of Auburn University must register and pay fees.

T F 82. Auburn University may release a student's educational records to representatives of the Secretary of H. E. W. without prior written consent.

T F (83.) Students are in the best position to observe and evaluate a professor's teaching performance.

APPENDIX C (Cont'd)

T F 84. At Auburn University a student with 140 quarter hours earned credit is considered a Senior.

T F 85. Critical self-analysis of performance greatly reduces the possibilities of learning.

T F 86. Auburn University has a Water Resources Research Institute.

T F 87. The Dean of Student Affairs of Auburn University has the responsibility for determining whether a student shall be classified as an Alabama or non-Alabama resident student.

T F 88. At Auburn University, students with a minimum overall grade average of 2.2 are graduated with Honor.

T F (89.) Faculty evaluation forms are the major criteria considered when faculty members are given tenure.

T F (90.) When rating a professor, you should compare his performance to how well you think you could do in his place.

T F 91. At Auburn University, if the instructor does not appear within 20 minutes after the hour, it may be assumed that the class is cancelled.

T F 92. At least two units of college preparatory mathematics are required for admission to any curriculum at Auburn University.

T F (93.) The "halo error" in rating refers to the practice of rating everyone high on the scale.

T F 94. At Auburn University, excuses for the purpose of attending reserve military training are normally denied.

T F 95. Verbal guidance and cues typically confuse the learner.

T F 96. Auburn University's fees are somewhat higher than those charged by similar institutions in the Southeast.

T F 97. Auburn University is an equal opportunity educational institution.

T F (98.) Professors teaching 500 and 600 level courses should not be rated any differently than those teaching 200 and 300 level courses.

T F (99.) Research and consulting are the Auburn University professors' most important duties.

T F 100. In a learning situation, simple tasks should be practiced as a whole.

APPENDIX C (Cont'd)

T F (101.) The use of systematic faculty evaluation procedures is intended to insure that all faculty members are treated fairly.

T F 102. Auburn University may deny admission to any individual whose presence is deemed detrimental to the institution or its students.

T F 103. Student discipline at Auburn University is under the supervision of the Dean of Student Affairs.

T F 104. Students have the right to review any financial records their parents submit to Auburn University.

T F 105. S and U grades do not enter into grade-point average computations at Auburn University.

T F 106. At Auburn University, arrangements to make up work missed due to absence should be initiated by the student.

T F 107. Auburn University's $10 admissions application processing fee is refundable on request.

T F 108. Auburn University lists conducting basic and applied research among its major purposes.

T F 109. Auburn University has a Cyber 70 computer as its primary computer.

T F 110. Auburn University at Montgomery offers bachelors and masters degrees.

T F 111. Auburn University offers doctorate degrees in 48 academic fields.

T F (112.) Faculty evaluation systems are intended mainly to identify and remove inferior teachers.

T F (113.) Part-time instructors and graduate students should be rated more leniently than full-time faculty members.

T F 114. Auburn University requires a minimum of 19 high school units for admission.

T F (115.) Faculty evaluation data play a part when decisions are made regarding professors' salaries and promotions.

T F (116.) When rating a professor's teaching effectiveness, it is important to base the ratings on one or two specific instances of extremely good or bad teaching behavior.

T F 117. "Directory Information" may not be released by Auburn University without the student's written consent.

APPENDIX C (Cont'd)

T F 118. To earn the bachelor's degree at Auburn University a student must earn at least a D average on credits accepted for his degree program.

T F 119. The Ralph Brown Draughan Library has more than 1,000,000 bound volumes in its current holdings.

T F 120. The Auburn campus is just a little over 500 acres in area.

T F 121. At Auburn University a student with 50 quarter hours earned credit is considered a Sophomore.

T F 122. The Auburn University Computer Center offers a masters degree in Computer Science.

T F 123. Auburn University inspects and approves suitable off-campus housing for its students.

T F (124.) How well you are doing in the course should not affect the ratings you give your professor.

T F 125. On approval of his dean, an Auburn University student may schedule up to 23 hours per quarter.

T F (126.) When rating a professor, you should compare his performance to that of the best teacher you ever had.

T F (127.) The fairest way for a department head to evaluate the performance of his faculty is to get the comments of two or three students majoring in his field.

T F (128.) Standardized teacher rating forms are intended to reduce bias in the performance evaluation process.

T F 129. Auburn University has a chartered Phi Beta Kappa chapter.

T F 130. There are over 150,000 Auburn alumni.

T F 131. "Overlearning" should be avoided in a learning situation.

T F 132. Students are more likely to learn an assignment if they see a purpose in what they are doing.

T F 133. Alabama residents are required to take the Scholastic Aptitude Test (SAT) before being considered for admission to Auburn University.

T F 134. Auburn University may place an undergraduate student on probation or suspension at any time if he flagrantly neglects his academic work or makes unsatisfactory progress toward graduation.

APPENDIX C (Cont'd)

T F 135. A student in good standing in an accredited college may be admitted to Auburn University as a transient student when faculty and facilities are available.

T F 136. Knowledge of the value of a task affects efforts to learn the task.

T F 137. Credit earned at another institution by a student on academic suspension from Auburn University can be used in meeting requirements for an Auburn University degree.

T F (138.) The primary use of faculty evaluation data is to influence decisions about promotions and salary.

T F 139. There are currently over 20,000 students enrolled at Auburn University.

T F 140. The student, in registering at Auburn Univeristy, agrees to conform with its rules and regulations for conduct.

T F (141.) You should typically try to give higher ratings to full professors than to associate or assistant professors.

T F 142. The use of rewards increases the possibility of learning.

T F 143. Repeated failure in a learning situation can lead to feelings of inadequacy.

T F 144. At Auburn University, the first ten days of each quarter are designed as the Special Examination period to remove X grades.

T F 145. Applications to Auburn University from out-of-state residents are accepted for all curricula except Pre-Veterinary Medicine.

T F 146. Auburn University was originally named East Alabama Male College.

T F 147. Attempts to recall learned material usually interfere with the learning process.

T F (148.) The best approach to evaluating teachers is to rate them highly unless they are obviously incompetent.

T F (149.) When rating professors, you should not let later information interfere with your first impressions.

T F (150.) Other professors are typically in the best position to evaluate a professor's classroom teaching performance.

APPENDIX D

Scale 5

What is your opinion of student evaluation of faculty teaching performance through the use of rating forms? (Check one of the eleven responses below.)

_____ 1. I am very strongly opposed to it.

_____ 2. I am strongly opposed to it.

_____ 3. I am opposed to it.

_____ 4. I am slightly opposed to it.

_____ 5. I am very slightly opposed to it.

_____ 6. I am neither opposed nor in favor of it.

_____ 7. I am very slightly in favor of it.

_____ 8. I am slightly in favor of it.

_____ 9. I am in favor of it.

_____ 10. I am strongly in favor of it.

_____ 11. I am very strongly in favor of it.

APPENDIX E-1

Dimension of Teacher Performance: _____

1. Describe an incident which you have observed which is an example of good teacher performance in this dimension. Be as specific as possible.

2. Describe an incident which you have observed which is an example of average teacher performance in this dimension. Be as specific as possible.

3. Describe an incident which you have observed which is an example of poor teacher performance in this dimension. Be as specific as possible.

APPENDIX E-2

ADOLESCENT BEHAVIOR DESCRIPTION QUESTIONNAIRE, FORM A

1. Describe an incident which happened to you during your adolescence which made you feel really good. Be as specific as possible.

2. Describe another incident which happened to you during your adolescence which made you feel really good. Be as specific as possible.

3. Describe another incident which happened to you during your adolescence which made you feel really good. Be as specific as possible.

APPENDIX E-3

ADOLESCENT BEHAVIOR DESCRIPTION QUESTIONNAIRE, FORM B

1. Describe an incident which happened to you during your adolescence which made you feel really bad. Be as specific as possible.

2. Describe another incident which happened to you during your adolescence which made you feel really bad. Be as specific as possible.

3. Describe another incident which happened to you during your adolescence which made you feel really bad. Be as specific as possible.

APPENDIX F-1

"Behaviorally-Anchored" Rating Scales for Evaluating
the Teaching Performance of College Professors

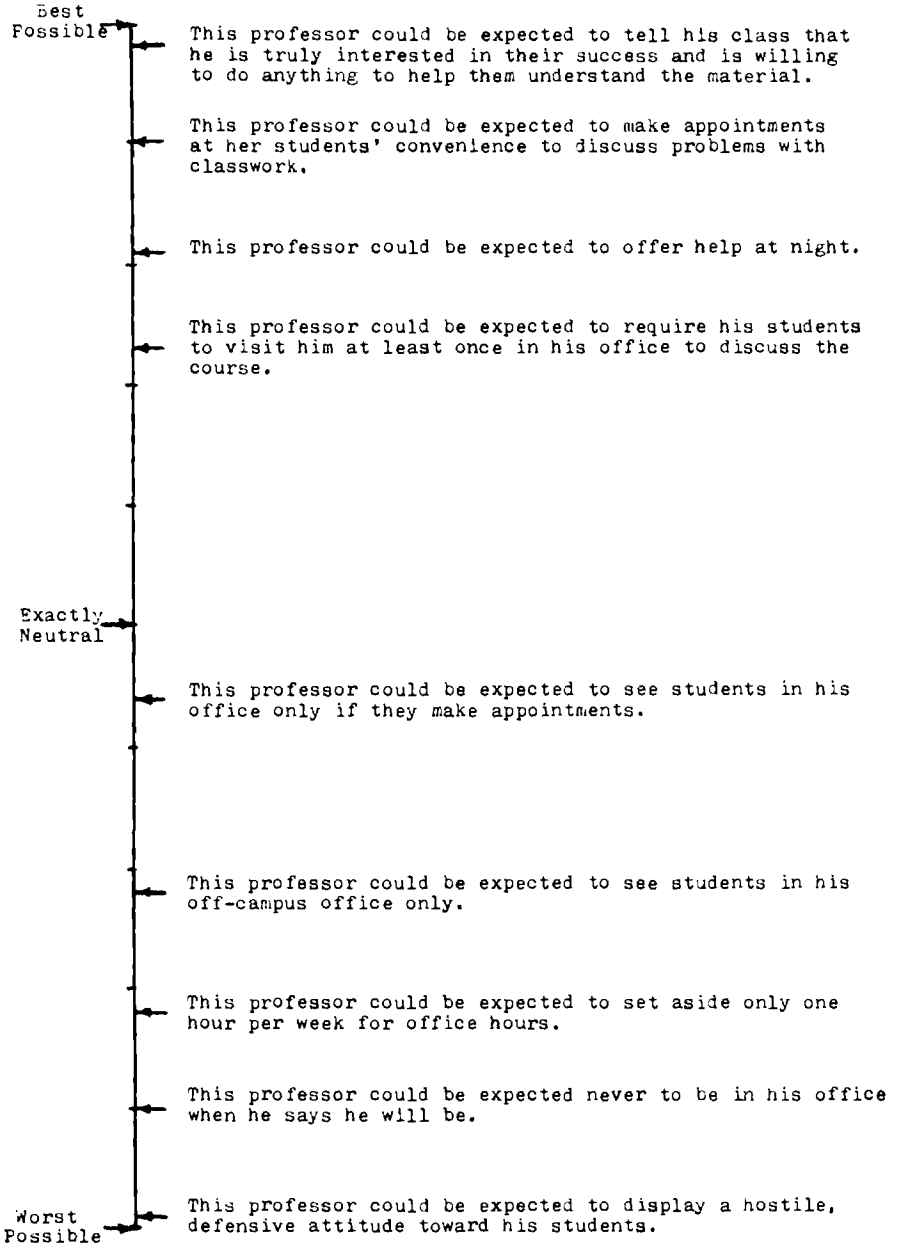
Attached are scales for evaluating the teaching performance of college professors on five dimensions: (A) Relationships with Students, (B) Ability to Present the Material, (C) Interest in Course and Material, (D) Reasonableness of the Workload, and (E) Fairness of Testing and Grading. (The scales in this packet have been randomly ordered so they may not follow this alphabetical arrangement.) You will note that below each dimension description is a vertical line divided into 10 segments. The lowest point on the line is labeled "worst possible" (performance), the highest point is labeled "best possible" (performance), and the midpoint is labeled "exactly neutral" (performance). Instead of being anchored by numbers or letters, these scales have descriptions of actual professors' behaviors to give you an idea of what each point on the scale means. Note that the arrows associated with the statements to the right of the vertical line point to the level of performance on the scale described by each statement. You are to evaluate your professor by placing an X on each vertical line at the point where you think his or her performance falls on the scale. For each dimension, read the description and all of the anchor statements, decide about where on the line your professor's level of performance would fall, and mark a neat X at that point. You can place your X anywhere along the line, not just at the points where segment marks or arrows are. Mark one and only one X on each of the five vertical lines.

Note. The scales have been reduced considerably in size. The vertical lines are 11 inches long in the actual BARS.

APPENDIX F-1 (Cont'd)

A. Relationships with Students

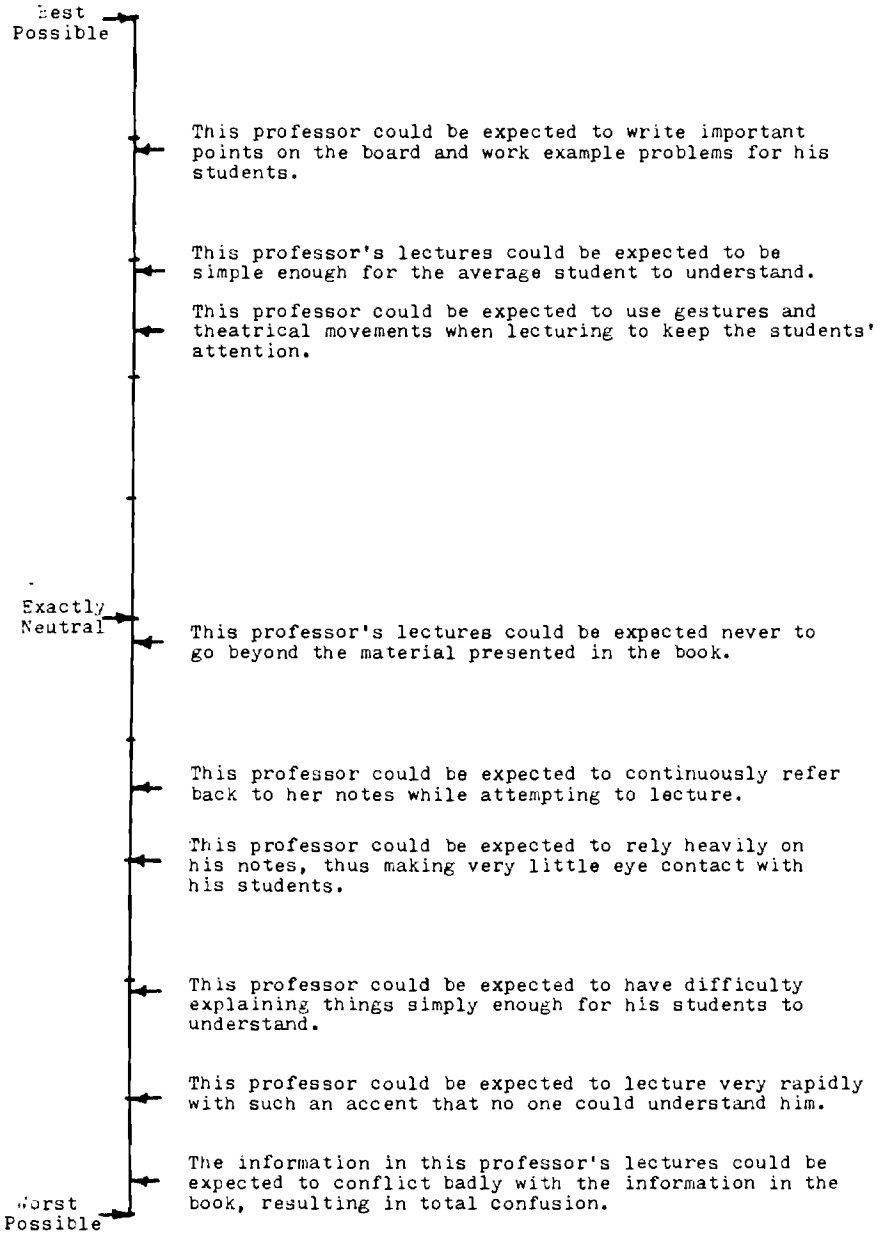
This dimension refers to the way the professor treats his/her students both in and out of class. It includes such things as talking with students before, during, and after class, interacting with and counseling students in the office and elsewhere regarding course-related and personal problems, knowing students' names, and treating students with respect in class.



APPENDIX F-1 (Cont'd)

B. Ability to Present the Material

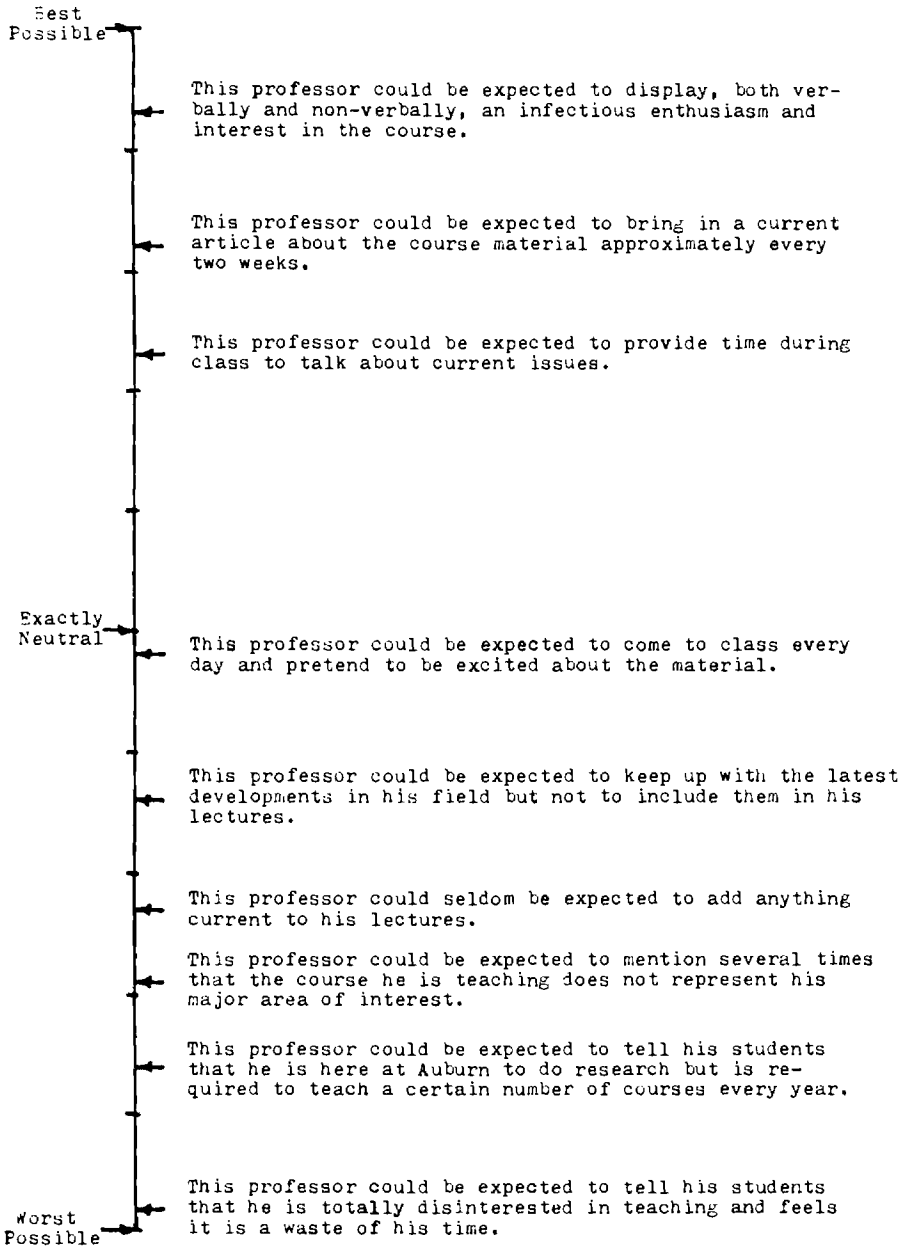
This dimension refers to the way the professor organizes the material and presents it to the class. It includes such things as coming to class well-prepared and on time, organizing the material in a logical manner, speaking and writing clearly, and using examples, audio-visual aids, and other devices to get the material across to the students.



APPENDIX F-1 (Cont'd)

C. Interest in Course and Material

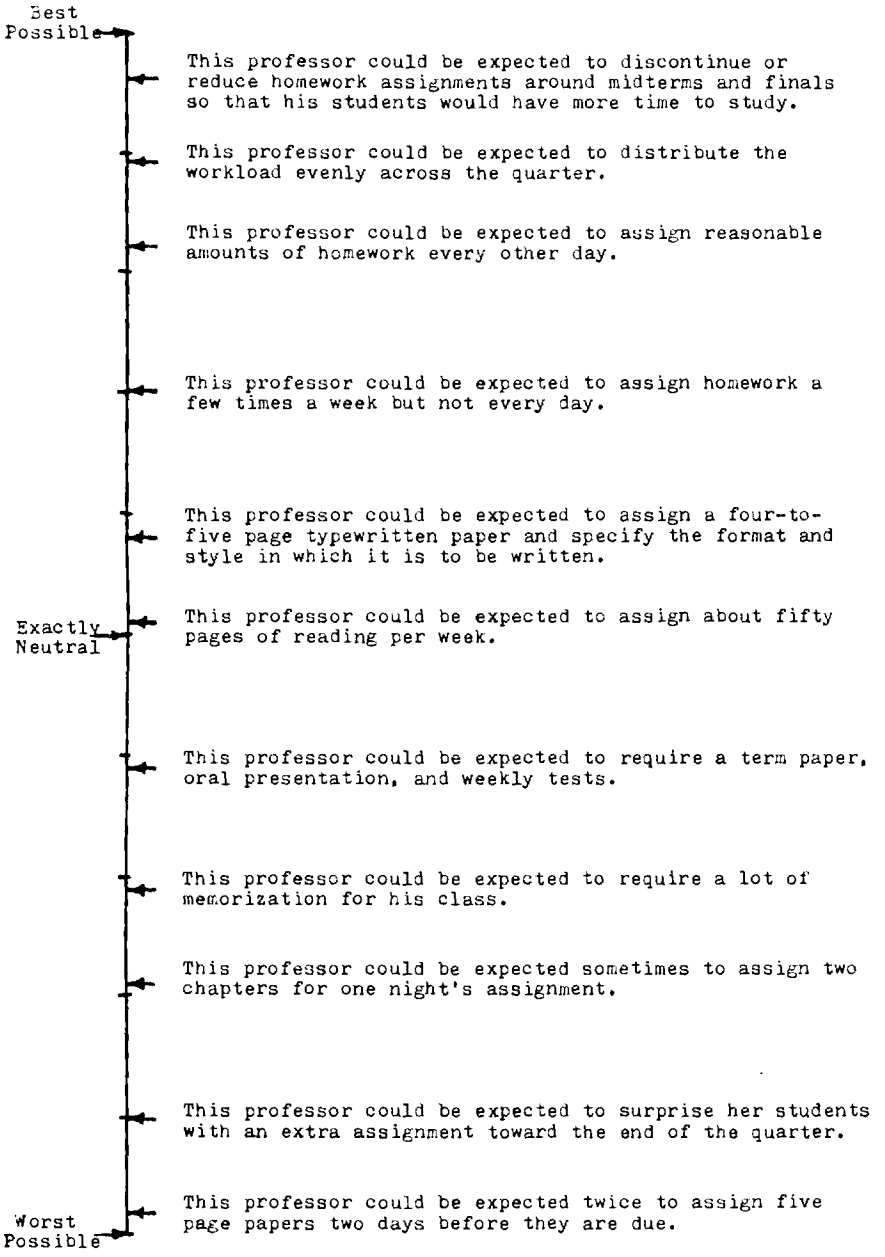
This dimension refers to the professor's knowledge of and interest in the material he/she is trying to teach. It includes such things as being able to answer questions and elaborate on the material, showing enthusiasm for the course, and reading and researching to keep current and learn more about the subject matter.



APPENDIX F-1 (Cont'd)

D. Reasonableness of the Workload

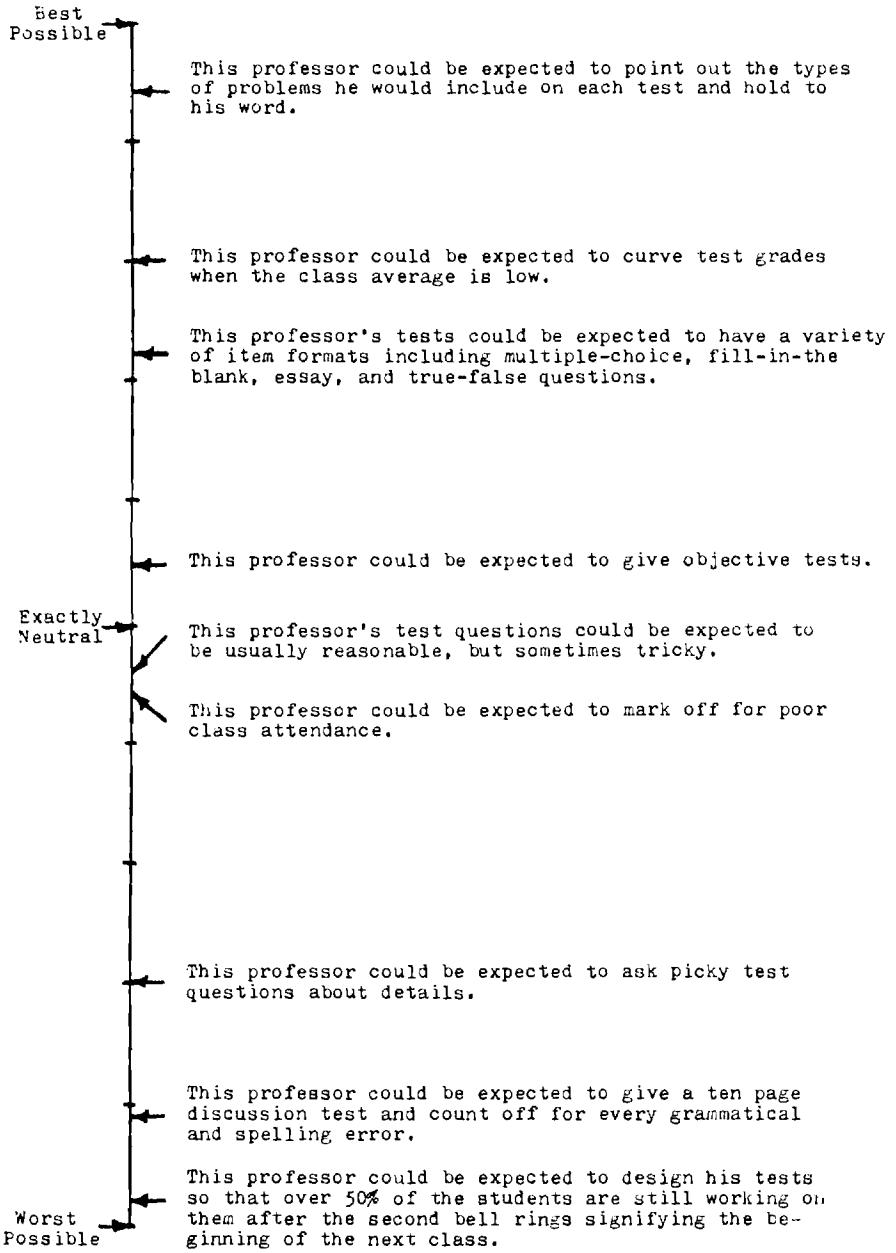
This dimension refers to the amount of work (reading, homework problems, class and lab work, papers, tests, etc.) assigned by the professor. It includes such things as clearly specifying assignments and due dates, scheduling the work evenly throughout the quarter, and keeping the workload appropriate to the credit-hour value of the course.



APPENDIX F-1 (Cont'd)

E. Fairness of Testing and Grading

This dimension refers to the fairness of the professor's testing and grading policies. It includes such things as stating how grades are to be determined, testing over appropriate material, and grading without bias.



APPENDIX F-2

BARS ITEM STATISTICS

Dimension	Percent Agreement in Category Placement	S Value	Q Value
A	100	10.8	1.0
A	96	10.0	1.4
A	96	9.1	1.8
A	100	8.3	1.5
A	100	5.4	1.9
A	100	3.8	2.0
A	100	2.8	1.3
A	96	2.0	1.3
A	100	1.1	0.6
B	96	9.9	1.3
B	88	8.9	1.5
B	84	8.4	1.6
B	60	5.8	2.1
B	100	4.6	2.2
B	92	4.0	1.8
B	100	2.9	1.5
B	100	2.0	1.3
B	88	1.3	1.2
C	96	10.3	1.4
C	92	9.2	1.4
C	68	8.3	2.0
C	96	5.8	2.5
C	80	4.6	2.4
C	96	3.7	1.5
C	100	3.1	2.0
C	100	2.4	1.6
C	92	1.2	0.7

APPENDIX F-2 (Cont'd)

Dimension	Percent Agreement in Category Placement	S Value	Q Value
D	92	10.6	1.3
D	92	9.9	1.9
D	100	9.2	1.6
D	96	8.0	1.9
D	88	6.8	1.3
D	100	6.1	2.1
D	96	4.9	2.5
D	92	3.9	2.1
D	96	3.1	1.4
D	84	2.0	0.9
D	100	1.2	0.9
E	100	10.4	1.3
E	100	9.0	1.8
E	100	8.2	2.8
E	88	6.5	2.4
E	96	5.6	1.9
E	72	5.4	3.0
E	100	3.0	1.8
E	100	1.9	1.4
E	100	1.2	0.9

APPENDIX G-1

PROFESSOR L

Professor L is a 29-year-old male Assistant Professor who is new at Auburn. He has long red hair, a full beard and moustache, and is a heavy smoker. He usually wears jeans and flannel shirts, boots, and a black leather jacket to class. He is not very well known in his field but has initiated a number of research projects since arriving at Auburn. He teaches a 5-hour, 300-level science course with a laboratory.

You observed the following things about Professor L while taking his course:

He used a variety of methods to present the material, including films, tapes, and experiments.

He told the class he would grade on a 10-point scale, then actually used a 7-point scale to assign final grades.

He often described his own fascination with the material he was covering.

He gave a mid-term and final only.

He assigned only as much homework as was necessary to learn the material thoroughly.

He was attentive and helpful in class, but was generally unavailable for outside help.

He gave plenty of time to read the material and discussed it thoroughly in class.

Once when asked a question in class he lost patience with himself because he could not answer it.

He always left promptly after giving his lectures.

When asked by his students what to study for a test, he said, "I don't know, I haven't made it out yet."

He did not curve grades even if the average score was in the 50s or 60s.

He gave a student unclear and evasive answers to her questions when she visited his office.

His lectures were boring and unorganized.

He assigned about two hours worth of work to be done during his three-hour laboratory so that no one would have to rush.

APPENDIX G-1 (Cont'd)

He took his lectures straight from the book and never gave examples.

He often told the class about interesting articles he had read or experiments he had heard about.

Although he gave his office number and hours on the first day of class, he did not encourage the students to come see him.

Once when confounded by a student's question in class he spent several hours of his own time that afternoon researching material for an answer.

He reduced the workload at the end of the quarter when he realized that his students did not have enough time to complete all of the assignments.

He sought student input to support his conclusions in class.

APPENDIX G-1 (Cont'd)

PROFESSOR M

Professor M is a 60-year-old female Associate Professor who has been teaching at Auburn off and on for over 30 years. She is tall and gray-haired, wears wire-rimmed glasses, and chain-smokes. She wears expensive suits and dresses and always appears well-groomed. Although she never earned a Ph. D., she is very well respected on campus for her work in faculty committees and community outreach programs. She teaches a 5-hour quantitative course with a laboratory devoted to problem-solving.

You observed the following things about Professor M while taking her course:

She often left out steps while working problems on the board and was unable to tell the students how she reached the solutions.

She assigned ten pages of reading each night.

Her lectures never seemed to have anything to do with the subject matter of the course.

She sought out a shy student who was failing and worked with her until she understood the material well enough to pass the course.

She never brought in outside material relating to the course.

She encouraged students to come to her office for help.

She used flowery language and talked above the heads of her students.

She gave deadlines for papers and dates of tests but changed them as the quarter progressed.

Her tests usually covered three or four chapters of the book.

She gave 13 lab quizzes and dropped the lowest three.

She assigned no more than two chapters of reading per week.

She allowed her students to call her at home if they had a problem and could not reach her at her office.

Her tests had a lot of questions so that you could miss one and not worry about failing.

APPENDIX G-1 (Cont'd)

She came to class and said, "Well, here we are, so I might as well lecture on something."

She gave positive feedback for responding to her questions in class even when the responses were not exactly correct.

She did not really understand the material she was presenting and ended up confusing the class.

She specified the exact chapters that would be covered on each test.

She acted as though it hurt her to teach class.

She assigned homework every night and checked it every Friday.

She told her class that she hates the textbook.

APPENDIX G-1 (Cont'd)

PROFESSOR N

Professor N is a 34-year-old male Associate Professor who has been at Auburn for six years. He is tall and somewhat heavy, with tousled black hair. He wears large dark-rimmed glasses and is always dressed in a dark, conservative suit, white shirt, and solid tie. He is very well respected as a writer and researcher and makes a great deal of money doing outside consulting. He teaches a four-hour, 500-level lecture course.

You observed the following things about Professor N while taking his course:

He assigned and tested over 5-8 chapters per week.

He would not assign work for several days, then would give a heavy assignment for a single night.

He noticed puzzled looks on his students' faces while he was lecturing and reworded his presentation so that they could understand.

He posted office hours but made his students wait until he could find time to see them.

He helped a student get through a personal crisis.

He assigned five 7-10 page reports within a four-week period, in addition to weekly tests and an average of 25 pages of reading per night.

His tests were ambiguous and much too long.

He gave hard tests which required the students to study a lot.

He never came to class unprepared.

He tried to relate complex material to the students in a manner that they could understand.

He always made up his tests a half hour before he gave them.

When making course assignments, he did not consider that students were taking courses other than his.

He brought in various films, magazines, and pictures to illustrate his lectures.

He compensated for limited office hours by offering his time before and after class every day.

APPENDIX G-1 (Cont'd)

He did not curve grades unless the class did extremely badly.

He failed to follow up on his promise to find out answers to questions asked in class.

He traveled in order to see and hear things about his profession which he then shared with his students.

He was willing to help students with special problems, whether personal or otherwise.

He brought in up-to-date material and gave the students interesting tid-bits related to the subject.

He frequently missed class and sent a graduate student in his place.

APPENDIX G-1 (Cont'd)

PROFESSOR O

Professor O is a 23-year-old female Graduate Teaching Assistant who has been at Auburn for one year. She has long, medium-brown hair and is rather attractive although somewhat thin. She usually wears stylish suits or dresses to class, although occasionally she dresses less formally. She is an outspoken advocate of liberal causes and is active in student government. She teaches a three-hour, 200-level literature course.

You observed the following things about Professor O while taking her course:

One day her classroom presentation consisted of reading, in a low monotone, the topical sentences from each paragraph of an out-of-date textbook.

She refused to set office hours.

She always acted excited and happy to be in class.

She refused to discuss grades in class.

On the first day of class, she told her class how interesting she found the subject and assured them that they would too.

She added points to her students' test grades if she found questions on them that no one answered correctly.

She took into consideration students' other classes and outside activities when assigning work.

She made her students feel uncomfortable as though they were wasting her time.

She gave a number of tests each of which covered a small amount of material.

She constantly criticized students' thoughts, ideas, and interpretations of material.

She rarely assigned homework.

She never brought her notes to class.

She gave short reading assignments.

She seemed to know something about all the different topics covered in the course.

She assigned either one chapter or two essays (never both) to be read each week.

APPENDIX G-1 (Cont'd)

She did not lecture, she just told her students to read the textbook and ask questions if they did not understand something.

She acted a little crazy at times to keep the class awake.

She cursed at her students for not commenting in class or reading assigned chapters.

She specified one thing which she said would be on the test, went over it in class before the test, then did not include it on the test.

She knew the material so well that she was able to answer all questions asked by her students.

APPENDIX G-1 (Cont'd)

PROFESSOR P

Professor P is a 57-year-old male full Professor who has been teaching at Auburn for 18 years. He is short and thin with graying hair and a somewhat unattractive face due to a scar on his left cheek. He typically wears business suits to class. He was once very important in his field, but his research activities have dwindled and his research projects are sometimes described as "dated". He taught a five-hour, 100-level laboratory science class.

You observed the following things about Professor P while taking his course:

He told his students how much each test and project was worth toward the final grade.

He assigned homework as a punishment and never graded it.

He did not know how to do some of the things he was supposed to be teaching his students to do.

He never changed his tone or expression while lecturing.

He announced his office hours so that students could see him if they needed to.

Instead of assigning homework, he told the students to work the problems they wanted to.

His tests covered only what he told his students would be on them.

He brought in new material to class to substitute for out-of-date, unclear material in the textbook.

He often could not answer questions because he had not even read the material he had assigned.

He tried to establish a feeling of equality between himself and his students.

He presented information in brief, easy-to-follow written outline form.

Whenever the answer to a test question was unclear, he always gave the benefit of the doubt to the student.

He always pointed out the most important aspects of the material covered for a test, then made sure the test questions came from the important material.

APPENDIX G-1 (Cont'd)

He gave an extremely heavy assignment one week, then slacked off for a week or so before giving another assignment.

He gave his students his office number but did not make them feel welcome.

He belittled the class material and described the course as a waste of time.

He required a typewritten lab report every week in addition to the regular course work.

He used good teaching aids, was articulate, and stressed important points in class.

He acted so bored with the material that he seemed almost to put himself to sleep.

He would answer questions only after class.

APPENDIX G-2

ITEM STATISTICS FOR SIMULATED PROFESSORS

Order in Diary	Dimension	Percentage of Agreement in Category Placement	S Value	Q Value
		Professor L		
12	A	72	2.7	1.7
17	A	84	4.1	2.9
6	A	92	4.6	2.4
9	A	68	4.6	2.7
			$\bar{x}$=4.0	
13	B	92	2.1	1.2
15	B	76	3.9	2.7
20	B	68	7.9	2.7
1	B	92	10.1	1.7
			$\bar{x}$=6.0	
8	C	68	3.4	2.6
3	C	84	9.0	2.2
16	C	80	9.4	2.6
18	C	84	10.3	1.3
			$\bar{x}$=8.0	
19	D	96	9.4	2.0
14	D	88	9.7	1.9
5	D	92	9.7	2.2
7	D	84	9.7	2.4
			$\bar{x}$=9.6	
2	E	92	1.4	1.2
11	E	96	2.0	1.6
4	E	96	2.0	1.8
10	E	76	2.6	2.1
			$\bar{x}$=2.0	

APPENDIX G-2 (Cont'd)

Order in Diary	Dimension	Percentage of Agreement in Category Placement	S Value	Q Value
		Professor M		
15	A	88	9.4	2.0
6	A	100	9.6	1.8
12	A	100	10.4	1.2
4	A	96	10.6	1.3
			$\bar{x}$=10.0	
3	B	88	1.7	1.3
16	B	80	2.0	1.1
1	B	92	2.0	1.7
7	B	88	2.3	1.5
			$\bar{x}$=2.0	
18	C	68	3.1	1.6
14	C	72	3.6	2.0
5	C	64	3.8	3.0
20	C	92	5.5	2.9
			$\bar{x}$=4.0	
8	D	76	4.8	2.8
19	D	92	5.9	2.1
2	D	100	6.1	2.4
11	D	100	7.2	3.0
			$\bar{x}$=6.0	
9	E	72	6.5	2.4
10	E	76	7.8	2.8
13	E	100	8.0	2.9
17	E	76	9.7	1.9
			$\bar{x}$=8.0	

APPENDIX G-2 (Cont'd)

Order in Diary	Dimension	Percentage of Agreement in Category Placement	S Value	Q Value
		Professor N		
4	A	96	3.1	2.0
14	A	100	8.4	2.6
5	A	92	10.1	1.1
18	A	100	10.4	1.2
			$\bar{x}$=8.0	
10	B	92	9.4	2.0
13	B	88	9.8	1.7
3	B	92	10.4	1.1
9	B	68	10.4	1.7
			$\bar{x}$=10.0	
20	C	76	2.2	1.3
16	C	68	2.9	1.8
19	C	76	9.2	1.4
17	C	96	9.7	1.4
			$\bar{x}$=6.0	
12	D	92	1.8	1.3
6	D	84	1.8	1.3
2	D	100	2.2	1.7
1	D	88	2.2	1.7
			$\bar{x}$=2.0	
7	E	92	2.4	1.5
11	E	88	2.4	2.0
8	E	76	5.6	1.9
15	E	92	5.6	3.0
			$\bar{x}$=4.0	

APPENDIX G-2 (Cont'd)

Order in Diary	Dimension	Percentage of Agreement in Category Placement	S Value	Q Value
		Professor O		
2	A	96	1.6	1.3
18	A	96	1.9	1.6
8	A	80	2.1	1.3
10	A	92	2.4	1.5
			$\bar{x}$=2.0	
1	B	84	1.6	1.2
16	B	60	1.9	1.3
12	B	84	4.9	3.0
17	B	60	7.6	2.3
			$\bar{x}$=4.0	
5	C	100	9.7	1.9
3	C	76	10.0	2.0
14	C	80	10.1	1.9
20	C	80	10.2	1.3
			$\bar{x}$=10.0	
11	D	100	6.3	2.9
13	D	92	8.1	2.4
15	D	96	8.4	1.6
7	D	84	9.2	2.2
			$\bar{x}$=8.0	
19	E	92	3.1	3.0
4	E	68	3.3	3.0
9	E	76	8.4	2.4
6	E	96	9.4	2.2
			$\bar{x}$=6.1	

APPENDIX G-2 (Cont'd)

Order in Diary	Dimension	Percentage of Agreement in Category Placement	S Value	Q Value
		Professor P		
20	A	72	2.8	1.5
15	A	100	3.1	2.0
5	A	96	8.7	1.9
10	A	100	9.4	1.8
			$\bar{x}$=6.0	
4	B	100	3.7	1.8
11	B	92	8.4	3.0
8	B	68	9.9	1.3
18	B	72	10.0	1.4
			$\bar{x}$=8.0	
16	C	92	1.2	0.8
9	C	68	1.9	1.0
19	C	84	2.1	1.7
3	C	68	2.8	1.8
			$\bar{x}$=2.0	
2	D	60	2.1	1.8
17	D	96	3.5	1.3
6	D	80	4.8	2.9
14	D	96	5.6	2.8
			$\bar{x}$=4.0	
12	E	92	9.6	2.7
7	E	96	10.0	1.8
1	E	80	10.0	2.2
13	E	100	10.4	1.2
			$\bar{x}$=10.0	

APPENDIX H

Rater Training Program

I. Clarification of the aims and purposes of rating.

A. The evaluation of professors' teaching performance is a commonly occurring event.

1. When we think of "faculty evaluation," we usually visualize a formal process involving rating forms, computer printouts, etc. Actually, the evaluation of professors' teaching performance occurs quite often, usually in an informal manner.

2. Students frequently "compare notes" and "spread the word" about professors. As they do this, the students are informally evaluating their professors, often on the basis of reputation and randomly observed events.

3. Professors often evaluate themselves and other professors in informal discussions. These informal evaluations also may be largely based on reputation and randomly observed events, as well as comments from two or three students.

4. Deans, department heads, and others are faced with making decisions regarding promotion, tenure, salary, course assignments, etc. for their professors. These decisions require some type of evaluation of the professors in question. When objective data are not available, these decisions are frequently based upon some type of informal evaluation, such as reputation, random observations, or the comments of two or three students or faculty members, even though these are certainly not the fairest ways to evaluate faculty members.

B. There is a need for systematic, objective information regarding teaching performance.

1. For lack of more objective data, important decisions are often made on the basis of the "informal evaluation" described above. As noted, much of this informal evaluation is based on hearsay, reputation, random comments and observations, etc. These sources are often inaccurate and even unfair. They typically present a distorted, biased picture of the professor's true teaching ability and performance. In order to increase the possibilities of appropriate, unbiased, fair decisions being made, it is necessary

APPENDIX H (Cont'd)

to gather more objective, systematic, relevant information about faculty teaching performance. Teacher rating forms are one means of making faculty evaluation more objective and systematic, and less biased.

2. One major problem with many faculty rating forms is that they can be interpreted differently by each student rater. Thus, characteristics of the type of form used can influence the outcome of a teacher evaluation project. Students do not always agree on the definition of "good teaching performance," and what one sees as "excellent" performance may be only "fair" to another. Since the outcome of the rating process can be as easily influenced by how the raters interpret the form as by the faculty member's actual teaching performance, it is important to make sure that all of the raters interpret the form as similarly as possible. The teaching behaviors to be evaluated and the meaning of each point on the scale should be clearly specified to ensure nearly uniform interpretation. Otherwise, the raters may all be rating different aspects of behavior, and the data will not be meaningful.

3. In order to be useful, faculty evaluation data must be reliable. That is, the evaluations by several independent raters of the same professor's teaching performance in the same class should be relatively consistent--there should be relatively high agreement among the raters. If there is a very low rate of agreement among the raters, the information will obviously be of little use.

C. Some uses of objective faculty evaluation data.

1. Feedback--Objective faculty evaluation data serve as relatively effective feedback from the students to their professors. Teacher evaluation forms enable students to communicate ideas to their teachers, to make their teachers aware of particular strengths and weaknesses in their courses and in their teaching methods, and to suggest improvements when necessary. Since learning depends on feedback, this information is essential if professors are to improve their courses and teaching methods in the future. The primary use of faculty rating forms is to provide this important feedback to the individual faculty members.

2. Personnel actions--Objective faculty evaluation data, when available, can be used to influence decisions regarding such issues as tenure, promotion, and

APPENDIX H (Cont'd)

salary adjustment. Decisions based on objective data are typically fairer than those based on hearsay, reputation, and other "informal" data. Student evaluations of teaching performance are rarely the major criteria considered when personnel action decisions are made, but they can certainly have some influence.

3. Development--Objective faculty evaluation data can help deans and department heads identify any particular training needs or special talents in their professors, thus providing them with suggestions for faculty development. Individual professors can also identify their own particular weaknesses and seek to improve themselves.

4. Placement--Objective faculty evaluation data can be used to influence decisions regarding course assignments, class sizes, etc.

5. Responsibility--The faculty evaluation process often enhances a professor's feelings of responsibility toward his/her students and duties as a teacher.

6. Effectiveness--Through the above uses, objective faculty evaluation data can help improve departmental, school, and university effectiveness, as well as the effectiveness of the individual faculty member.

D. Additional points regarding faculty evaluation.

1. There are many different duties involved in the job of college professor. While classroom teaching is not the professor's only responsibility, it is an important part of his/her job. Auburn University lists teaching as its faculty members' most important duty.

2. Students are not the only persons whose evaluations of teaching performance should be sought, but their evaluations should be considered carefully. Students are one of the major consumer groups of the university's expertise and are certainly affected by the faculty's performance. Furthermore, whereas deans, department heads, and other faculty members rarely observe professors' teaching performance first-hand, and thus are not in a strong position to provide objective data, students are in an excellent position to observe and report on faculty teaching behavior.

APPENDIX H (Cont'd)

3. Teaching is multi-dimensional. There are many facets of teaching performance and it is probably not possible to take all of them into account in any one performance measure or rating form. The rating form should, however, cover as many important teaching behaviors as possible and should certainly provide adequate coverage of the facets it is intended to measure.

4. The purpose of faculty evaluation is to improve professors' teaching performance, not to damage faculty members in any way. The process should only be used constructively, never destructively.

II. Introduction of the Behaviorally Anchored Rating Scale

A. Most faculty rating forms are developed by administrators or faculty committees with limited student input. The scale used in this project, however, was developed through student participation, and is intended to be clear and meaningful to student raters. The scale dimensions and behavioral examples were provided by Auburn students participating in earlier phases of this study.

(Note: At this point in the training session the scale is shown to the trainees. A full description of the scale includes the following points.)

B. Instruction on the meaning of the characteristics to be evaluated.

C. Instruction on the meaning of each anchor point used on the scale.

D. Instruction on how to use the scale.

III. Instruction on the avoidance of common pitfalls in rating.

A. Lack of objectivity.

1. Some student raters evaluate their professors on the basis of supposition, guesswork, and reputation, thus defeating the entire purpose of using the rating forms. A student's rating of his/her professor should be based only upon first-hand observations of actual behaviors, not comments made by other students, reputational factors, etc. A student who has not observed a teacher first-hand should not evaluate that teacher. Nor should a student let his/her rating of a professor be influenced by what other students think.

APPENDIX H (Cont'd)

2. Some student raters base their entire rating of a professor on one or two instances of extremely good or extremely poor teaching behavior. While these isolated extreme instances should certainly be considered, it is important also to keep in mind the typical, "day-in, day-out" behavior of the professor.

3. All students tend to have "first impressions" of their teachers, but some students never change these impressions, even in the face of behaviors to the contrary, and base their ratings exclusively on their first impressions. The professor's behavior throughout the quarter should be considered when his/her performance is being evaluated.

4. The most common problem involving lack of objectivity is allowing some biasing factor to affect a professor's rating. As difficult as it is, student raters should strive not to let such factors as the professor's age, sex, rank, or appearance, the course's level or difficulty, or the student's own performance (i.e., grade in the course) or personal liking or disliking of the professor influence the performance ratings given to the professor. A student's rating of his/her professor should be influenced <u>only</u> by the professor's actual behavior while teaching the course, <u>not</u> by any biasing factor. Non-teaching behaviors, such as consulting and research, should also typically be ignored when the professor's <u>teaching</u> performance is being evaluated.

B. Common rating "errors" to avoid.

(<u>Note</u>: This presentation is accompanied by a visual display of how these errors would appear on the Behaviorally Anchored Rating Scale.)

1. <u>Leniency</u>--This "error" occurs when the student rates the professor (and probably other professors) higher on every item of the rating scale than the professor's true level of performance actually deserves.

2. <u>Severity</u>--This "error," the opposite of leniency, occurs when the student rates the professor (and probably other professors) lower on every item of the rating scale than the professor's true level of performance actually deserves.

3. <u>Central tendency</u>--This "error" occurs when the student uses only the central portion of the scale, ignoring the high and low extremes, even when the

APPENDIX H (Cont'd)

the professor's true level of performance deserves an unusually high or low rating.

4. Extremity--This "error," the opposite of central tendency, occurs when the student uses only the high and low extremes of the scale, ignoring the central portion, even when the professor's true level of performance deserves a more moderate rating.

5. Halo--This "error" occurs when the student forms a general, overall impression of the professor's performance, then fills out the rating form to reflect this impression. This practice should be avoided. Instead, the student rater should consider each item on the scale individually, and should try not to let his/her rating of the teacher on one item influence the rating on another item.

6. Logical--This "error," similar to the "halo error," occurs when the student, in an attempt to appear consistent, bases his/her rating on "logic" rather than observation, thus allowing his/her response to one scale item to unjustly influence the response to another. As stated above, each item on the scale should be considered individually. A professor's level of performance will typically not be perfectly consistent (from item to item), thus there is no requirement that the students' rating of the professor be somehow logically consistent. What is important is that the ratings reflect only the professor's actual level of performance on each item.

7. Proximity--This "error," similar to the two above, occurs when the student allows his/her rating on one item of the scale to influence the rating on a second item simply because the two items are located close to one another on the scale. Again, each item should be considered independently.

8. Contrast and comparison--These "errors" occur when the student rates his professor not according to the standards specified on the scale, but in contrast or comparison to some other kind of standard, such as the performance of the best or worst professor the student has ever known, the level of performance the student thinks he/she could attain if he/she were teaching the course, etc. Each professor should be evaluated independently according to the standards specified on the rating scale, not in comparison with other teachers, ideals, etc.

APPENDIX H (Cont'd)

IV. Practice in the use of the scales.

(Note: During the time remaining in the training session, students practice using the Behaviorally Anchored Rating Scales to evaluate professors of their own choosing. [No professors are identified.] Students are encouraged to examine their own ratings for examples of bias and error, and to correct their ratings when appropriate.)

BIBLIOGRAPHY

Reference Notes

1. Ronan, W. W. Criterion contamination and performance. Unpublished manuscript, 1974. (Available from W. W. Ronan, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332).

2. Ronan, W. W. Performance behavior and criteria. Unpublished manuscript, 1974. (Available from W. W. Ronan, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332).

3. Tauscher, L. J. Individual perspectives towards dimensions of college teaching effectiveness. Unpublished master's thesis, Georgia Institute of Technology, 1975.

4. Levy, S., & Stone, D. M. Process and content of managerial ratings of subordinates. Paper presented at the meeting of the Eastern Psychological Association, New York, April 1963.

5. Arauz, C. G. The effects of office noise upon decisions made in a personnel manager simulation. Unpublished master's thesis, Georgia Institute of Technology, 1975.

6. Ronan, W. W. Definition of performance criteria. Unpublished manuscript, 1974. (Available from W. W. Ronan, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia 30332).

7. Rotter, N., & Rotter, G. S. Race, work performance and merit rating: An experimental evaluation. Paper presented at the meeting of the Eastern Psychological Association, Philadelphia, April 1969.

8. Lopez, F. M., Jr. The blood, sweat and tears of employee performance evaluation. Speech delivered at the annual conference of the Public Personnel Association, October 1966.

9. Bernardin, H. J., & Boetcher, R. The effects of rater training and cognitive complexity on psychometric error in ratings. Manuscript submitted for publication, 1978.

10. Wakeley, J. H. The effects of specific training on accuracy in judging others. Unpublished doctoral dissertation, Michigan State University, 1961.

11. Pond, S. B., III, & Sauser, W. I., Jr. A longitudinal investigation of the effects of rater training and rater participation in scale construction on a measure of cognitive complexity. Manuscript in preparation, 1978.

12. Bernardin, H. J. Personal communication, March, 1978.

References

American Psychological Association. Standards for educational & psychological tests. Washington, D.C.: Author, 1974.

Anderson, S.B., Ball, S., Murphy, R. T., and Associates. Encyclopedia of educational evaluation. San Francisco: Jossey-Bass, 1975.

Arvey, R. D., & Hoyle, J. C. A Guttman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 1974, 59, 61-68.

Astin, A. W. Criterion-centered research. Educational and Psychological Measurement, 1964, 24, 807-822.

Atlanta Regional Commission. Test validation project report: Volume 1 (with appendices). Atlanta: Author, 1974.

Ayers, A. W. A comparison of certain visual factors with the efficiency of textile inspectors. Journal of Applied Psychology, 1942, 26, 812-827.

Baridon, F. E., & Loomis, E. H. Personnel problems: Methods of analysis and control. New York: McGraw-Hill, 1931.

Barr, A. J., & Goodnight, J. H. SAS: A user's guide to the Statistical Analysis System. Raleigh, N.C.: North Carolina State University, 1972.

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. A user's guide to SAS 76. Raleigh, N.C.: SAS Institute, 1976.

Barrett, R. S. Performance rating. Chicago: Science Research Associates, 1966.

Barrett, R. S., Taylor, E. K., Parker, J. W., & Martens, L. Rating scale content: 1. Scale information and supervisory ratings. Personnel Psychology, 1958, 11, 333-346.

Bartlett, F. C. Remembering: A study in experimental and social psychology. New York: Macmillan, 1932.

Bass, B. M. The leaderless group discussion as a leadership evaluation instrument. Personnel Psychology, 1954, 7, 470-477.

Bass, B. M. Reducing leniency in merit ratings. Personnel Psychology, 1956, 9, 359-369.

Bass, B. M. Further evidence on the dynamic character of criteria. Personnel Psychology, 1962, 15, 93-97.

Bass, B. M., & Barrett, G. V. Man, work, and organizations: An introduction to industrial and organizational psychology. Boston: Allyn and Bacon, 1972.

Bavelas, A., & Strauss, G. Group dynamics and intergroup relations. In W. G. Gennis, K. D. Benne, & R. Chin (Eds.), The planning of change. New York: Holt, 1961.

Baylie, T. N., Kujawski, C. J., & Young, D. M. Appraisals of "people" resources. In D. Yoder & H. G. Heneman, Jr. (Eds.), ASPA handbook of personnel and industrial relations: Vol. 1. Staffing policies and strategies. Washington, D.C.: Bureau of National Affairs, 1974.

Bayroff, A. G., & Burke, J. H. The rater's guide. Personnel Psychology, 1950, 3, 461-465.

Bayroff, A. G., Haggerty, H. R., & Rundquist, E. A. Validity of ratings as related to rating techniques and conditions. Personnel Psychology, 1954, 7, 93-113.

Bechtoldt, H. P. Problems in establishing criterion measures. In D. B. Stuit (Ed.), Personnel research and test development in the Bureau of Naval Personnel. Princeton, N.J.: Princeton University Press, 1947.

Behrend, H. Absence and labour turnover in a changing economic climate. Occupational Psychology, 1953, 27, 69-79.

Bellows, R. M. Procedures for evaluating vocational criteria. Journal of Applied Psychology, 1941, 25, 499-513.

Bellows, R. M. Psychology of personnel in business and industry (3rd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1961.

Benjamin, R., Jr. A survey of 130 merit-rating plans. Personnel, 1952, 29, 289-294.

Bentz, V. J. The Sears experience in the investigation, description, and prediction of executive behavior. In Predicting managerial success. Ann Arbor, Mich.: Foundation for Research in Human Behavior, 1968.

Berkowitz, L. The judgmental process in personality functioning. Psychological Review, 1960, 67, 130-142.

Berkshire, J. R., & Highland, R. W. Forced-choice performance rating--A methodological study. Personnel Psychology, 1953, 6, 355-378.

Bernardin, H. J. Behavioral expectation scales versus summated scales: A fairer comparison. Journal of Applied Psychology, 1977, 62, 422-427.

Bernardin, H. J. Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 1978, 63, 301-308.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. A recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 1976, 61, 564-570.

Bernardin, H. J., LaShells, M. B., Smith, P. C., & Alvares, K. M. Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 1976, 61, 75-79.

Bernardin, H. J., & Walter, C. S. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Besco, R. O., & Lawshe, C. H. Foreman leadership as perceived by superiors and subordinates. Personnel Psychology, 1959, 12, 573-582.

Besnard, G. C., & Briggs, L. J. Measuring job proficiency by means of a performance test. In E. A. Fleishman (Ed.), Studies in personnel and industrial psychology (Rev. ed.). Homewood, Ill.: Dorsey Press, 1967.

Bigoness, W. J. Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. Journal of Applied Psychology, 1976, 61, 80-84.

Bingham, W. V. Halo, invalid and valid. Journal of Applied Psychology, 1939, 23, 221-228.

Bingham, W. V., & Davis, W. T. Intelligence test scores and business success. Journal of Applied Psychology, 1924, 8, 1-22.

Bittner, R. H. Developing an industrial merit rating procedure. Personnel Psychology, 1948, 1, 403-432.

Bittner, R. H., & Rundquist, E. A. The rank-comparison rating method. Journal of Applied Psychology, 1950, 34, 171-177.

Blau, P. M. The dynamics of bureaucracy: A study of interpersonal relations in two government agencies (Rev. ed.). Chicago: University of Chicago Press, 1963.

Blood, M. R. Spin-offs from behavioral expectation scale procedures. Journal of Applied Psychology, 1974, 59, 513-515.

Blum, M. L., & Naylor, J. C. Industrial psychology: Its theoretical and social foundations. New York: Harper & Row, 1968.

Blumberg, H. H., DeSoto, C. B., & Kuethe, J. L. Evaluation of rating scale formats. Personnel Psychology, 1966, 19, 243-259.

Borman, W. C. The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 1974, 12, 105-124.

Borman, W. C. Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 1975, 60, 556-560.

Borman, W. C. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, 63, 135-144.

Borman, W. C., & Dunnette, M. D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Borman, W. C., & Vallon, W. R. A view of what can happen when behavioral expectation scales are developed in one setting and used in another. Journal of Applied Psychology, 1974, 59, 197-201.

Boruch, R. F., Larkin, J. D., Wolins, L., & MacKinney, A. C. Alternative methods of analysis: Multitrait-multimethod data. Educational and Psychological Measurement, 1970, 30, 833-853.

Bray, D. W. The assessment center method of appraising management potential. In J. W. Blood (Ed.), The personnel job in a changing world. New York: American Management Association, 1964.

Bray, D. W., & Campbell, R. J. Selection of salesmen by means of an assessment center. Journal of Applied Psychology, 1968, 52, 36-41.

Bray, D. W., & Grant, D. L. The assessment center in the measurement of potential for business management. Psychological Monographs, 1966, 80 (17, Whole No. 625).

Brogden, H. E., & Taylor, E. K. The dollar criterion: Applying the cost accounting concept to criterion construction. Personnel Psychology, 1950, 3, 133-154. (a)

Brogden, H. E., & Taylor, E. K. The theory and classification of criterion bias. Educational and Psychological Measurement, 1950, 10, 159-186. (b)

Brown, E. M. Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 1968, 52, 195-199.

Bruner, J. S., & Tagiuri, R. The perception of people. In G. Lindzey (Ed.), Handbook of social psychology: Volume II. Special fields and applications. Cambridge: Addison-Wesley, 1954.

Buckner, D. N. The predictability of ratings as a function of interrater agreement. Journal of Applied Psychology, 1959, 43, 60-64.

Burke, R. J., & Goodale, J. G. New way to rate nurse performance. Hospitals, 1973, 47(24), 62-68.

Burnaska, R. F., & Hollmann, T. D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 1974, 59, 307-312.

Byham, W. C. Assessment centers for spotting future managers. Harvard Business Review, 1970, 48(4), 150-160, plus appendix.

Byham, W. C., & Thornton, G. C. III. Assessment centers: A new aid in management selection. Studies in Personnel Psychology, 1970, 2(2), 21-35.

Campbell, D. T., & Chapman, J. P. Testing for stimulus equivalence among authority figures by similarity in trait description. Journal of Consulting Psychology, 1957, 21, 253-256.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, J. P., Dunnette, M. D., Arvey, R. D., & Hellervik, L. V. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E. Managerial behavior, performance, and effectiveness. New York: McGraw-Hill, 1970.

Campion, J. E. Work sampling for personnel selection. Journal of Applied Psychology, 1972, 56, 40-44.

Carter, L. F., & Dudek, F. J. The use of psychological techniques in measuring and critically analyzing navigators' flight performance. Psychometrika, 1947, 12, 31-42.

Cascio, W. F., & Valenzi, E. R. Behaviorally anchored rating scales: Effects of education and job experience of raters and ratees. Journal of Applied Psychology, 1977, 62, 278-282.

Cascio, W. F., & Valenzi, E. R. Relations among criteria of police performance. Journal of Applied Psychology, 1978, 63, 22-28.

Champney, H. The measurement of parent behavior. Child Development, 1941, 12, 131-166.

Champney, H., & Marshall, H. Optimal refinement of the rating scale. Journal of Applied Psychology, 1939, 23, 323-331.

Charest, H. G., Cowart, D. G., & Goodman, P. S. Multi-instrument, multi-rater, multi-trait method for assessing measures of managerial performance. Experimental Publication System, 1963, 3, no. 092A.

Christal, R. E., & Madden, J. M. Effect of degree of familiarity in job evaluation. USAF WADD Personnel Laboratory Technical Note, November 1960, No. 60-263.

Cliff, N. Adverbs as multipliers. Psychological Review, 1959, 66, 27-44.

Coch, L., & French, J. R. P., Jr. Overcoming resistance to change. Human Relations, 1948, 1, 512-532.

Cohen, J., & Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1975.

Conrad, H. S. The bogey of the "personal equation" in ratings of intelligence. Journal of Educational Psychology, 1932, 23, 147-149. (a)

Conrad, H. S. The personal equation in ratings: I. An experimental determination. Journal of Genetic Psychology, 1932, 41, 267-293. (b)

Conrad, H. S. The personal equation in ratings: II. A systematic evaluation. Journal of Educational Psychology, 1933, 24, 39-46.

Cox, J. A., & Krumboltz, J. D. Racial bias in peer ratings of basic airmen. Sociometry, 1958, 21, 292-299.

Crawford, P. S., & Bradshaw, H. L. Perception of characteristics of effective university teachers: A scaling analysis. Educational and Psychological Measurement, 1968, 28, 1079-1085.

Creswell, M. B. Effects of confidentiality on performance ratings of professional health personnel. Personnel Psychology, 1963, 16, 385-393.

Cronbach, L. J. Processes affecting scores on "understanding of others" and "assumed similarity." Psychological Bulletin, 1955, 52, 177-193.

Cronbach, L. J. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Daniels, H. W., & Edgerton, H. A. The development of criteria of safe operation for groups. Journal of Applied Psychology, 1954, 38, 47-53.

Davis, N. M. A study of a merit-rating scheme in a factory. Occupational Psychology, 1953, 27, 57-68.

Deaux, J. E., & Emswiller, T. Explanations of successful performance on sex-linked tasks: What is skill for the male is luck for the female. Journal of Personality and Social Psychology, 1947, 29, 80-85.

Deaux, K., & Taynor, J. Evaluation of male and female ability: Bias works two ways. Psychological Reports, 1973, 32, 261-262.

deJung, J. E., & Kaplan, H. Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. Journal of Applied Psychology, 1962, 46, 370-374.

Dickinson, T. L., & Tice, T. E. A multitrait-multimethod analysis of scales developed by retranslation. Organizational Behavior and Human Performance, 1973, 9, 421-438.

Dickinson, T. L., & Tice, T. E. The discriminant validity of scales developed by retranslation. Personnel Psychology, 1977, 30, 217-228.

Division of Industrial-Organizational Psychology, American Psychological Association. Principles for the validation and use of personnel selection procedures. Dayton, Ohio: The Industrial-Organizational Psychologist, 1975.

Dorcus, R. M. Methods of evaluating the efficiency of door-to-door salesmen of bakery products. Journal of Applied Psychology, 1940, 24, 587-594.

Draper, R. D. Comparison of promotional judgments by superiors, peers, and subordinates. American Psychologist, 1964, 19, 449.

Dunnette, M. D. A note on the criterion. Journal of Applied Psychology, 1963, 47, 251-254.

Dunnette, M. D. Personnel selection and placement. Belmont, Calif.: Brooks/Cole, 1966.

Dyer, R., Matthews, J. J., Stulac, J. F., Wright, C. E., & Yudowitch, K. Questionnaire construction literature survey: Final report, annex 2 (Contract No. DAHC19-74-C-0032). Fort Hood, Tex.: Army Research Institute for Behavioral and Social Sciences, June 1975.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Edwards, A. L. Techniques of attitude scale construction. New York: Appleton-Century-Crofts, 1957.

Einhorn, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. Organizational Behavior and Human Performance, 1971, 6, 1-27.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. Adoption by four agencies of uniform guidelines on employee selection procedures (1978). Federal Register, 1978, 43, 38289-38315.

Ewart, E. S., Seashore, S. E., & Tiffin, J. A factor analysis of an industrial merit rating scale. Journal of Applied Psychology, 1941, 25, 481-486.

Famularo, J. J. (Ed.). Handbook of modern personnel administration. New York: McGraw-Hill, 1972.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Ethnic group membership as a moderator of the prediction of job performance. Personnel Psychology, 1971, 24, 609-636.

Fechner, G. T. Elemente der psychophysik. Leipzig: Breitkopf and Hartel, 1860.

Ferguson, L. W. The development of a method of appraisal for assistant managers. Journal of Applied Psychology, 1947, 31, 306-311.

Ferguson, L. W. The value of acquaintance ratings in criterion research. Personnel Psychology, 1949, 2, 93-102. (a)

Ferguson, L. W. The effect upon appraisal scores of individual differences in the ability of superiors to appraise subordinates. Personnel Psychology, 1949, 2, 377-382. (b)

Festinger, L. A theory of cognitive dissonance. Evanston, Ill.: Row, Peterson, 1957.

Finkle, R. B. Managerial assessment centers. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Fiske, D. W. Values, theory, and the criterion problem. Personnel Psychology, 1951, 4, 93-98.

Flanagan, J. C. (Ed.). The aviation psychology program in the Army Air Forces: Report no. 1. Washington, D.C.: U.S. Government Printing Office, 1948.

Flanagan, J. C. Critical requirements: A new approach to employee evaluation. Personnel Psychology, 1949, 2, 419-425.

Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Flanagan, J. C., & Burns, R. K. The employee performance record: A new appraisal and development tool. Harvard Business Review, 1955, 5, 95-102.

Flanagan, J. C., Fiske, D. W., Bass, B. M., Carter, L. F., Kelly, E. L., & Weislogel, R. L. Situational performance tests: A symposium. Personnel Psychology, 1954, 7, 461-498.

Flaugher, R. L., Campbell, J. T., & Pike, L. W. Prediction of job performance for Negro and white medical technicians: Ethnic group membership as a moderator of supervisor's ratings (PR-69-5). Princeton, N.J.: Educational Testing Service, 1969.

Fleishman, E. A. Attitude versus skill factors in work group productivity. Personnel Psychology, 1965, 18, 253-266.

Fleishman, E. A., & Fruchter, B. Factor structure and predictability of successive stages of learning Morse code. Journal of Applied Psychology, 1960, 44, 97-101.

Fleishman, E. A., & Hempel, W. E., Jr. Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 1954, 19, 239-252.

Fogli, L., Hulin, C. L., & Blood, M. R. Development of first-level behavioral job criteria. Journal of Applied Psychology, 1971, 55, 3-8.

Frederiksen, N., Saunders, D. R., & Wand, B. The in-basket test. Psychological Monographs, 1957, 71(9, Whole No. 438).

French, J. R. P., Jr., Israel, J., & As, D. An experiment on participation in a Norwegian factory. Human Relations, 1960, 13, 3-19.

French, J. R. P., Jr., Kay, E., & Meyer, H. H. Participation and the appraisal system. Human Relations, 1966, 19, 3-20.

Freyd, M. The graphic rating scale. Journal of Educational Psychology, 1923, 14, 83-102.

Friedlander, F. Performance and interactional dimensions of organizational work groups. Journal of Applied Psychology, 1966, 50, 257-265.

Friedman, B. A., & Cornelius, E. T., III. Effect of rater participation in scale construction on the psychometric characteristics of two rating scale formats. Journal of Applied Psychology, 1976, 61, 210-216.

Gage, N. L., & Cronbach, L. J. Conceptual and methodological problems in interpersonal perception. Psychological Review, 1955, 62, 411-422.

Galton, F. Inquiries into human faculty and its development. London: Macmillan, 1883.

Garner, W. R. Rating scales, discriminability, and information transmission. Psychological Review, 1960, 67, 343-352.

Gaylord, R. H., Russell, E., Johnson, C., & Severin, D. The relation of ratings to production records: An empirical study. Personnel Psychology, 1951, 4, 363-371.

Ghiselli, E. E. Dimensional problems of criteria. Journal of Applied Psychology, 1956, 40, 1-4.

Ghiselli, E. E., & Brown, C. W. Personnel and industrial psychology (2nd ed.). New York: McGraw-Hill, 1955.

Ghiselli, E. E., & Haire, M. The validation of selection tests in the light of the dynamic character of criteria. Personnel Psychology, 1960, 13, 225-231.

Gifford, W. S. Does business want scholars? Harper's Magazine, 1928, 156, 669-674.

Glickman, A. S. Effects of negatively skewed ratings on motivations of the rated. Personnel Psychology, 1955, 8, 39-47.

Goldberg, P. Are women prejudiced against men? Trans-action, 1968, 5(5), 28-30.

Goodale, J. G., & Burke, R. J. Behaviorally based rating scales need not be job specific. Journal of Applied Psychology, 1975, 60, 389-391.

Gordon, M. E. The effect of the correctness of the behavior observed on the accuracy of ratings. Organizational Behavior and Human Performance, 1970, 5, 366-377.

Graham, C. H. Behavior, perception and the psychophysical methods. Psychological Review, 1950, 57, 108-120.

Grant, D. L. A factor analysis of managers' ratings. Journal of Applied Psychology, 1955, 39, 283-286.

Greenhaus, J. F., & Gavin, J. F. The relationship between expectancies and job behavior for white and black employees. Personnel Psychology, 1972, 25, 449-455.

Grey, R. J., & Kipnis, D. Untangling the performance appraisal dilemma: The influence of perceived organizational context on evaluative processes. Journal of Applied Psychology, 1976, 61, 329-335.

Griffitt, W. Environmental effects on interpersonal affective behavior: Ambient effective temperature and attraction. Journal of Personality and Social Psychology, 1970, 15, 240-244.

Griffitt, W., & Veitch, R. Hot and crowded: Influences of population density and temperature on interpersonal behavior. Journal of Personality and Social Psychology, 1971, 17, 92-98.

Gruenfeld, L. W., & Weissenberg, P. Relationship between supervisory cognitive style and social orientation. Journal of Applied Psychology, 1974, 59, 386-388.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1936.

Guilford, J. P. Psychometric methods (2nd ed.). New York: McGraw-Hill, 1954.

Guilford, J. P. Fundamental statistics in psychology and education (4th ed.). New York: McGraw-Hill, 1965.

Guilford, J. P., Christensen, P. R., Taaffe, G., & Wilson, R. C. Ratings should be scrutinized. Educational and Psychological Measurement, 1962, 22, 439-447.

Guilford, J. P., & Fruchter, B. Fundamental statistics in psychology and education (6th ed.). New York: McGraw-Hill, 1978.

Guion, R. M. Criterion measurement and personnel judgments. Personnel Psychology, 1961, 14, 141-149.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Gulliksen, H. Foreword to W. S. Torgerson, Theory and methods of scaling. New York: Wiley, 1958.

Hakel, M. D., Ohnesorge, J. P., & Dunnette, M. D. Interviewer evaluations of job applicants' resumes as a function of the qualifications of the immediately preceding applicants: An examination of contrast effects. Journal of Applied Psychology, 1970, 54, 27-30.

Halsey, G. D. Supervising people (Rev. ed.). New York: Harper Brothers, 1953.

Hamner, W. C., Kim, J. S., Baird, L., & Bigoness, W. J. Race and sex as determinants of ratings by potential employers in a simulated work-sampling task. Journal of Applied Psychology, 1974, 59, 705-711.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.

Hart, H., & Olander, E. Sex differences in character as indicated by teachers' ratings. School and Society, 1924, 20, 381-382.

Hartshorne, H., & May, M. A. Studies in service and self-control. New York: Macmillan, 1929.

Hausman, H. J., & Strupp, H. H. Non-technical factors in supervisors' ratings of job performance. Personnel Psychology, 1955, 8, 201-217.

Hay, E. N. Predicting success in machine bookkeeping. Journal of Applied Psychology, 1943, 27, 483-493.

Helson, H. Adaptation-level as a frame of reference for prediction of psychophysical data. American Journal of Psychology, 1947, 60, 1-30.

Helwig, J. T. (Ed.). SAS supplemental library user's guide. Raleigh, N.C.: SAS Institute, 1977.

Hemphill, J. K., & Sechrest, L. B. A comparison of three criteria of aircrew effectiveness in combat over Korea. Journal of Applied Psychology, 1952, 36, 323-327.

Henry, W. E. Executive personality and job success. AMA Personnel Series, 1948, No. 120.

Hollander, E. P. The friendship factor in peer nominations. Personnel Psychology, 1956, 9, 435-447.

Hollander, E. P. The reliability of peer nominations under various conditions of administration. Journal of Applied Psychology, 1957, 41, 85-90.

Hollingworth, H. L. Judgments of persuasiveness. Psychological Review, 1911, 18, 234-256.

Hollingworth, H. L. Judging human character. New York: Appleton-Century-Crofts, 1922.

Holmes, D. S., & Berkowitz, L. Some contrast effects in social perception. Journal of Abnormal and Social Psychology, 1961, 62, 150-152.

Hoyle, J. C., & Arvey, R. D. Development of behaviorally based rating scales. Proceedings of the 10th Annual Computer Personnel Research Conference, June 15-16, 1972, 10, 85-103.

Hulin, C. L. The measurement of executive success. Journal of Applied Psychology, 1962, 46, 303-306.

Hunt, R. G. Interpersonal strategies for system management: Applications of counseling and participative principles. Monterey, Calif.: Brooks/Cole, 1974.

Huse, E. F., & Taylor, E. K. Reliability of absence measures. Journal of Applied Psychology, 1962, 46, 159-160.

Jackson, D. N. Multimethod factor analysis in the evaluation of convergent and discriminant validity. Psychological Bulletin, 1969, 72, 30-49.

Jacobson, M. B., & Effertz, J. Sex roles and leadership perceptions of the leaders and the led. Organizational Behavior and Human Performance, 1974, 12, 383-396.

James, L. R. Criterion models and construct validity for criteria. Psychological Bulletin, 1973, 80, 75-83.

Jenkins, J. G. Validity for what? Journal of Consulting Psychology, 1946, 10, 93-98.

Johnson, D. M., & Vidulich, R. N. Experimental manipulation of the halo effect. Journal of Applied Psychology, 1956, 40, 130-134.

Jones, E. E., Kanouse, D. E., Kelley, H. H., Nisbett, R. E., Valins, S., & Weiner, B. (Eds.). Attribution: Perceiving the causes of behavior. Morristown, N.J.: General Learning Press, 1971.

Jones, F. N. Overview of psychophysical scaling methods. In E. C. Carterette & M. P. Friedman (Eds.), Handbook of perception: Volume II. Psychophysical judgment and measurement. New York: Academic Press, 1974.

Kafry, D., Zedeck, S., & Jacobs, R. The scalability of behavioral expectation scales as a function of developmental criteria. Journal of Applied Psychology, 1976, 61, 519-522.

Kane, J. S., & Lawler, E. E., III. Methods of peer assessment. Psychological Bulletin, 1978, 85, 555-586.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: Multitrait-multimethod analyses of ratings. Psychological Bulletin, 1971, 75, 34-49.

Kay, B. R. The use of critical incidents in a forced-choice scale. Journal of Applied Psychology, 1959, 43, 269-270.

Keaveny, T. J., & McGann, A. F. A comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 1975, 60, 695-703.

Keith, J. A. H. The mutual influence of feelings. Harvard Psychological Studies, 1906, 2, 141-157.

Kelley, H. H. The process of causal attribution. American Psychologist, 1973, 28, 107-128.

Kellogg, M. S. What to do about performance appraisal. New York: American Management Association, 1965.

Kerr, W. A., Koppelmeier, G. J., & Sullivan, J. J. Absenteeism, turnover, and morale in a metals fabrication factory. Occupational Psychology, 1951, 25, 50-55.

King, G. F., Ehrmann, J. C., & Johnson, D. M. Experimental analysis of the reliability of observations of social behavior. Journal of Social Psychology, 1952, 35, 151-160.

Kingsbury, F. A. Analyzing ratings and training raters. Journal of Personnel Research, 1922, 1, 377-383.

Kipnis, D. Some determinants of supervisory esteem. Personnel Psychology, 1960, 13, 377-391.

Kirchner, W. K. Predicting ratings of sales success with objective performance information. Journal of Applied Psychology, 1960, 44, 398-403.

Kirchner, W. K. Relationships between supervisory and subordinate ratings for technical personnel. Journal of Industrial Psychology, 1966, 3, 57-60.

Kirchner, W. K., & Reisberg, D. J. Differences between better and less-effective supervisors in appraisal of subordinates. Personnel Psychology, 1962, 15, 295-302.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, Calif.: Brooks/Cole, 1968.

Klemmer, E. T., & Lockhead, G. R. Productivity and errors in two keying tasks: A field study. Journal of Applied Psychology, 1962, 46, 401-408.

Klimoski, R. J., & London, M. Role of the rater in performance appraisal. Journal of Applied Psychology, 1974, 59, 445-451.

Klores, M. S. Rater bias in forced-distribution performance ratings. Personnel Psychology, 1966, 19, 411-421.

Knauft, E. B. Construction and use of weighted check-list rating scales for two industrial situations. Journal of Applied Psychology, 1948, 32, 63-70.

Koltuv, B. B. Some characteristics of intrajudge trait intercorrelations. Psychological Monographs, 1962, 76(33, Whole No. 552).

Kornhauser, A. W. What are rating scales good for? Journal of Personnel Research, 1926, 5, 189-193.

Kornhauser, A. W. A comparison of ratings on different traits. Journal of Personnel Research, 1927, 5, 440-446.

Krasner, L. Behavior therapy. Annual Review of Psychology, 1971, 22, 483-532.

Landy, F. J., Farr, J. L., Saal, F. E., & Freytag, W. R. Behaviorally anchored scales for rating the performance of police officers. Journal of Applied Psychology, 1976, 61, 750-758.

Landy, F. J., & Guion, R. M. Development of scales for the measurement of work motivation. Organizational Behavior and Human Performance, 1970, 5, 93-103.

Latham, G. P., & Wexley, K. N. Behavioral observation scales for performance appraisal purposes. Personnel Psychology, 1977, 30, 255-268.

Latham, G. P., Wexley, K. N., & Pursell, E. D. Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 1975, 60, 550-555.

Lawler, E. E., III. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 369-381.

Lawshe, C. H., Kephart, N. C., & McCormick, E. J. The paired comparison technique for rating performance of industrial employees. Journal of Applied Psychology, 1949, 33, 69-77.

Lefkowitz, J. Effect of training on the productivity and tenure of sewing machine operators. Journal of Applied Psychology, 1970, 54, 81-86.

Levine, J., & Butler, J. Lecture vs. group decision in changing behavior. Journal of Applied Psychology, 1952, 36, 29-33.

Lifson, K. A. Errors in time-study judgments of industrial work pace. Psychological Monographs, 1953, 67(5, Whole No. 355).

Lindman, H. R. Analysis of variance in complex experimental designs. San Francisco: W. H. Freeman, 1974.

Lopez, F. M., Jr. Evaluating executive decision making: The in-basket technique. AMA Research Study 75. New York: American Management Association, 1966.

Lowin, A. Participative decision making: A model, literature critique, and prescriptions for research. Organizational Behavior and Human Performance, 1968, 3, 68-106.

Luck, T. J. Personnel audit and appraisal. New York: McGraw-Hill, 1955.

Lundy, T. M. Self-perceptions regarding masculinity-femininity and descriptions of same and opposite sex sociometric choices. Sociometry, 1958, 21, 238-246.

Lupton, T. On the shop floor. Oxford: Pergamon Press, 1963.

Maas, J. B. Patterned scaled expectation interview: Reliability studies on a new technique. Journal of Applied Psychology, 1965, 49, 431-433.

MacKinney, A. C., & Wolins, L. Validity information exchange no. 13-01. Personnel Psychology, 1960, 13, 443-447.

Madden, J. M. Familiarity effects in evaluative judgments. USAF WADD Personnel Laboratory Technical Note, November 1960, No. 60-261.

Madden, J. M. A further note on the familiarity effect in job evaluation. USAF ASD Personnel Laboratory Technical Note, June 1961, No. 61-47.

Madden, J. M. A comparison of three methods of rating-scale construction. Journal of Industrial Psychology, 1964, 2, 43-50.

Madden, J. M., & Bourdon, R. D. Effects of variations in rating scale format on judgment. Journal of Applied Psychology, 1964, 48, 147-151.

Mahler, W. R. Twenty years of merit rating: 1926-1946. New York: The Psychological Corporation, 1947.

Maier, N. R. F. Assets and liabilities in group problem solving: The need for an integrative function. Psychological Review, 1967, 74, 239-249.

Major, D. R. On the affective tone of simple sense-impressions. American Journal of Psychology, 1895, 7, 57-77.

Mandell, M. M. Supervisory characteristics and ratings. Personnel, 1956, 32, 435-440.

Marrow, A. J., Bowers, D. G., & Seashore, S. E. Management by participation. New York: Harper & Row, 1967.

Marsh, S. E., & Perrin, F. A. C. An experimental study of the rating scale technique. Journal of Abnormal and Social Psychology, 1925, 19, 383-399.

Martin, L. J. Psychology of aesthetics: I. Experimental prospecting in the field of the comic. American Journal of Psychology, 1905, 16, 35-118.

Maslow, A. H., & Zimmerman, W. College teaching ability, scholarly activity and personality. Journal of Educational Psychology, 1956, 47, 185-189.

McClelland, J. N., & Rhodes, F. Prediction of job success for hospital aides and orderlies from MMPI scores and personal history data. Journal of Applied Psychology, 1968, 53, 49-54.

McCormick, E. J., & Tiffin, J. Industrial psychology (6th ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1974.

Metzner, H., & Mann, F. Employee attitudes and absences. Personnel Psychology, 1953, 6, 467-485.

Meyer, H. H., Kay, E., & French, J. R. P., Jr. Split roles in performance appraisal. Harvard Business Review, 1964, 43, 124-129.

Mirvis, P. H., & Lawler, E. E., III. Measuring the financial impact of employee attitudes. Journal of Applied Psychology, 1977, 62, 1-8.

Mischel, W. Personality and assessment. New York: Wiley, 1968.

Mischel, W. Introduction to personality. New York: Holt, Rinehart and Winston, 1971.

Mitchell, T. R. Expectancy models of job satisfaction, occupational preference and effort: A theoretical, methodological, and empirical appraisal. Psychological Bulletin, 1974, 81, 1053-1077.

Morrison, R. F., Owens, W. A., Glennon, J. R., & Albright, L. E. Factored life history antecedents of industrial research performance. Journal of Applied Psychology, 1962, 46, 281-284.

Morse, N. C., & Reimer, E. The experimental change of a major organizational variable. Journal of Abnormal and Social Psychology, 1956, 52, 120-129.

Mosier, C. I. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. Psychological Review, 1940, 47, 355-366.

Motowidlo, S. J., & Borman, W. C. Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 1977, 62, 177-183.

Mueser, R. E. The weather and other factors influencing employee punctuality. Journal of Applied Psychology, 1953, 37, 329-337.

Mulaik, S. A. Are personality factors raters' conceptual factors? Journal of Consulting Psychology, 1964, 28, 506-511.

Mullins, C. J., & Force, R. C. Rater accuracy as a generalized ability. Journal of Applied Psychology, 1962, 46, 191-193.

Murray, H. A. Explorations in personality. New York: Oxford University Press, 1938.

Nagle, B. F. Criterion development. Personnel Psychology, 1953, 6, 271-289.

Newcomb, T. An experiment designed to test the validity of a rating technique. Journal of Educational Psychology, 1931, 22, 279-289.

Obradovic, J. Modification of the forced-choice method as a criterion of job proficiency. Journal of Applied Psychology, 1970, 54, 228-233.

Owens, W. A., Jr. Intra-individual differences versus inter-individual differences in motor skills. Educational and Psychological Measurement, 1942, 2, 209-314.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martens, L. Rating scale content: III. Relationship between supervisory- and self-ratings. Personnel Psychology, 1959, 12, 49-63.

Passini, F. T., & Norman, W. T. A universal conception of personality structure? Journal of Personality and Social Psychology, 1966, 4, 44-49.

Passini, F. T., & Norman, W. T. Ratee relevance in peer nominations. Journal of Applied Psychology, 1969, 53, 185-187.

Patchen, M. Participation, achievement, and involvement on the job. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Paterson, D. G. Methods of rating human qualities. Annals of the American Academy of Political and Social Science, 1923, 110, 81-93.

Pearson, K. On the relationship of intelligence to size and shape of head, and to other physical and mental characters. Biometrika, 1907, 5, 105-146.

Peters, D. L., & McCormick, E. J. Comparative reliability of numerically anchored versus job-task anchored rating scales. Journal of Applied Psychology, 1966, 50, 92-96.

Pheterson, G. I., Kiesler, S. B., & Goldberg, P. A. Evaluation of the performance of women as a function of their sex, achievement, and personal history. Journal of Personality and Social Psychology, 1971, 19, 114-118.

Prien, E. P. Assessments of high-level personnel: V. An analysis of interviewers' predictions of job performance. Personnel Psychology, 1962, 15, 319-334.

Prien, E. P. Dynamic character of criteria: Organizational change. Journal of Applied Psychology, 1966, 50, 501-504.

Rambo, W. W. The construction and analysis of a leadership behavior rating form. Journal of Applied Psychology, 1958, 42, 409-415.

Richardson, M. W. Forced-choice performance reports: A modern merit-rating method. Personnel, 1949, 26, 205-212.

Richardson, M. W., & Kuder, G. F. Making a rating scale that measures. Personnel Journal, 1933, 12, 36-40.

Roach, D. E., & Wherry, R. J., Sr. Performance dimensions of multi-line insurance agents. Personnel Psychology, 1970, 23, 239-250.

Ronan, W. W. A factor analysis of eight job performance measures. Journal of Industrial Psychology, 1963, 1, 107-112. (a)

Ronan, W. W. A factor analysis of eleven job performance measures. Personnel Psychology, 1963, 16, 255-267. (b)

Ronan, W. W. Evaluation of three criteria of management performance. Journal of Industrial Psychology, 1970, 5, 18-28.

Ronan, W. W. Development of an instrument to evaluate college classroom teaching effectiveness (Project No. 1-D-045; Grant No. OEG-4-71-0067). Washington, D.C.: U.S. Department of HEW, Office of Education, National Center for Educational Research and Development, 1971.

Ronan, W. W. Evaluating college classroom teaching effectiveness: PREP report no. 34. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1972.

Ronan, W. W., Anderson, C. L., & Talbert, T. L. A psychometric approach to job performance: Fire fighters. Public Personnel Management, 1976, 5, 409-422.

Ronan, W. W., & Latham, G. P. The reliability and validity of the critical incident technique: A closer look. Studies in Personnel Psychology, 1974, 6(1), 53-64.

Ronan, W. W., & Prien, E. P. Toward a criterion theory: A review and analysis of research and opinion. Greensboro, N.C.: The Richardson Foundation, 1966.

Ronan, W. W., & Schwartz, A. P. Ratings as performance criteria. International Review of Applied Psychology, 1974, 23, 71-82.

Rosen, B., & Jerdee, T. H. The influence of sex-role stereotypes on evaluations of male and female supervisory behavior. Journal of Applied Psychology, 1973, 57, 44-48.

Rosen, B., & Jerdee, T. H. Effects of applicant's sex and difficulty of job on evaluations of candidates for managerial positions. Journal of Applied Psychology, 1974, 59, 511-512.

Ross, P. F. Reference groups in man-to-man job performance rating. Personnel Psychology, 1966, 19, 115-142.

Rothe, H. F. Output rates among butter wrappers: I. Work curves and their stability. Journal of Applied Psychology, 1946, 30, 199-211. (a)

Rothe, H. F. Output rates among butter wrappers: II. Frequency distributions and an hypothesis regarding the "restriction of output." Journal of Applied Psychology, 1946, 30, 320-327. (b)

Rothe, H. F. Output rates among machine operators: I. Distributions and their reliability. Journal of Applied Psychology, 1947, 31, 484-489.

Rothe, H. F. The relation of merit ratings to length of service. Personnel Psychology, 1949, 2, 237-242.

Rothe, H. F. Output rates among chocolate dippers. Journal of Applied Psychology, 1951, 35, 94-97.

Rothe, H. F. Output rates among industrial employees. Journal of Applied Psychology, 1978, 63, 40-46.

Rothe, H. F., & Nye, C. T. Output rates among coil winders. Journal of Applied Psychology, 1958, 42, 182-186.

Rothe, H. F., & Nye, C. T. Output rates among machine operators: II. Consistency related to methods of pay. Journal of Applied Psychology, 1959, 43, 417-420.

Rothe, H. F., & Nye, C. T. Output rates among machine operators: III. A nonincentive situation in two levels of business activity. Journal of Applied Psychology, 1961, 45, 50-54.

Rotter, G. S., & Tinkleman, V. Anchor effects in the development of behavior rating scales. Educational and Psychological Measurement, 1970, 30, 311-318.

Rowe, P. M. Order effects in assessment decisions. Journal of Applied Psychology, 1967, 51, 170-173.

Rowland, V. K. Evaluating and improving managerial performance. New York: McGraw-Hill, 1970.

Rundquist, E. A., & Bittner, R. H. Using ratings to validate personnel instruments: A study in method. Personnel Psychology, 1948, 1, 163-183.

Rundquist, E. A., & Bittner, R. H. A merit rating procedure developed by and for the raters. Personnel, 1950, 26, 273-283.

Rush, C. H., Jr. A factorial study of sales criteria. Personnel Psychology, 1953, 6, 9-24.

Ryder, R. Teaching reliable rating of a process variable. Journal of Consulting Psychology, 1962, 26, 106.

Sauser, W. I., Jr., Arauz, C. G., & Chambers, R. M. Exploring the relationship between level of office noise and salary recommendations: A preliminary research note. Journal of Management, 1978, 4, 57-63.

Schmidt, F. L., & Johnson, R. H. The effect of race on peer ratings in an industrial situation. Journal of Applied Psychology, 1973, 57, 237-241.

Schmidt, F. R., & Kaplan, L. B. Composite versus multiple criteria: A review and resolution of the controversy. Personnel Psychology, 1971, 24, 419-434.

Schneider, B., & Bartlett, C. J. Industrial differences and organizational climate II.: Measurement of organizational climate by the multi-trait, multi-rater matrix. Personnel Psychology, 1970, 23, 493-512.

Schneider, D. E., & Bayroff, A. G. The relationship between rater characteristics and validity of ratings. Journal of Applied Psychology, 1953, 37, 278-280.

Schneier, C. E. Operational utility and psychometric characteristics of behavioral expectation scales: A cognitive reinterpretation. Journal of Applied Psychology, 1977, 62, 541-548.

Schneier, C. E., & Beatty, R. W. The influence of role prescriptions on the performance appraisal process. Academy of Management Journal, 1978, 21, 129-135.

Schuh, A. J. The predictability of employee tenure: A review of the literature. Personnel Psychology, 1967, 20, 133-152.

Schultz, D. G., & Siegel, A. I. Generalized Thurstone and Guttman scales for measuring technical skills in job performance. Journal of Applied Psychology, 1961, 45, 137-142.

Schwab, D. P., Heneman, H. G., & DeCotiis, T. A. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Scott, W. E., Jr., & Hamner, W. C. The influence of variations in performance profiles on the performance evaluation process: An examination of the validity of the criterion. Organizational Behavior and Human Performance, 1975, 14, 360-370.

Seashore, S. The aptitude hypothesis in motor skills. Journal of Experimental Psychology, 1931, 14, 555-561.

Seashore, S. E., Indik, B. P., & Georgopoulos, B. S. Relationships among criteria of job performance. Journal of Applied Psychology, 1960, 44, 195-202.

Selltiz, C., Wrightsman, L. S., & Cook, S. W. Research methods in social relations (3rd ed.). New York: Holt, Rinehart and Winston, 1976.

Severin, D. The predictability of various kinds of criteria. Personnel Psychology, 1952, 5, 93-104.

Sharon, A. T., & Bartlett, C. J. Effect of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology, 1969, 22, 251-263.

Shweder, R. A. How relevant is an individual difference theory of personality? Journal of Personality, 1975, 43, 453-484.

Sisson, E. D. Forced choice--The new Army rating. Personnel Psychology, 1948, 1, 365-381.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Smith, P. C., & Gold, R. A. Prediction of success from examination of performance during the training period. Journal of Applied Psychology, 1956, 40, 83-86.

Spicer, L. G. A survey of merit rating in industry. Personnel, 1951, 27, 515-518.

Spriegel, W. R., & Mumma, E. W. Merit rating of supervisors and executives. Austin, Tx.: Bureau of Business Research, University of Texas, 1961.

Springer, D. Ratings of candidates for promotion by co-workers and supervisors. Journal of Applied Psychology, 1953, 37, 347-351.

Stander, N. E. A longitudinal study of some relationships among criteria of managerial performance as perceived by superiors and subordinates. Journal of Industrial Psychology, 1965, 3, 43-51.

Stanley, J. C. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. Psychometrika, 1961, 26, 205-219.

Stark, S. Research criteria of executive success. Journal of Business, 1959, 32, 1-14.

Stevens, S. S. The psychophysics of sensory function. American Scientist, 1960, 48, 226-253.

Stockford, L., & Bissell, H. W. Factors involved in establishing a merit-rating scale. Personnel, 1949, 26, 94-118.

Stogdill, R. M., Shartle, C. L., Wherry, R. J., & Jaynes, W. E. A factorial study of administrative behavior. Personnel Psychology, 1955, 8, 165-180.

Strauss, G. Some notes on power-equalization. In H. Leavitt (Ed.), The social science of organizations: Four perspectives. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Symonds, P. M. On the loss of reliability in ratings due to coarseness of the scale. Journal of Experimental Psychology, 1924, 7, 456-461.

Symonds, P. M. Notes on rating. Journal of Applied Psychology, 1925, 9, 188-195.

Taft, R. The ability to judge people. Psychological Bulletin, 1955, 52, 1-23.

Talbert, T. L., Carroll, K. I., & Ronan, W. W. Measuring clerical job performance. Personnel Journal, 1976, 55, 573-575.

Tannenbaum, A. S. Social psychology of the work organization. Belmont, Calif.: Brooks/Cole, 1966.

Tate, B. L. A method for rating the proficiency of the hospital general staff nurse. New York: Research and Studies Service, National League for Nursing, 1964.

Taylor, C. W., Brice, P. B., Richards, J. M., Jr., & Jacobsen, T. L. An investigation of the criterion problem for a medical school faculty. Journal of Applied Psychology, 1964, 48, 294-301.

Taylor, C. W., Brice, P. B., Richards, J. M., Jr., & Jacobsen, T. L. An investigation of the criterion problem for a group of medical general practitioners. Journal of Applied Psychology, 1965, 49, 399-406.

Taylor, C. W., Smith, W. R., Ghiselin, B., & Ellison, R. Exploration in the measurement and prediction of contributions of one sample of scientists (Report No. ASD-RR-60-69). Lackland Air Force Base, Tex.: Personnel Laboratory, 1961.

Taylor, E. K., Barrett, R. S., Parker, J. W., & Martens, L. Rating scale content: II. Effect of rating on individual scales. Personnel Psychology, 1958, 11, 519-533.

Taylor, E. K., & Hastman, R. Relation of format and administration to the characteristics of graphic rating scales. Personnel Psychology, 1956, 9, 181-206.

Taylor, E. K., & Manson, G. E. Supervised ratings--Making graphic scales work. Personnel, 1951, 27, 504-514.

Taylor, E. K., Schneider, D. E., & Clay, H. Short forced-choice ratings work. Personnel Psychology, 1954, 7, 245-252.

Taylor, E. K., Schneider, D. E., & Symons, N. A. A short forced-choice evaluation form for salesmen. Personnel Psychology, 1953, 6, 393-401.

Taylor, E. K., & Wherry, R. J. A study of leniency in two rating systems. Personnel Psychology, 1951, 4, 39-47.

Taylor, J. G., & Smith, P. C. An investigation of the shape of learning curves for industrial motor tasks. Journal of Applied Psychology, 1956, 40, 142-149.

Thorndike, E. L. A constant error in psychological ratings. Journal of Applied Psychology, 1920, 4, 25-29.

Thorndike, R. L., & Hagen, E. Measurement and evaluation in psychology and education (3rd ed.). New York: John Wiley & Sons, 1969.

Thurstone, L. L. A law of comparative judgment. Psychological Review, 1927, 34, 273-286.

Thurstone, L. L., & Chave, E. J. The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church. Chicago: University of Chicago Press, 1929.

Tiffin, J., & Phelan, R. F. Use of the Kuder Preference Record to predict turnover in an industrial plant. Personnel Psychology, 1953, 6, 195-204.

Titchener, E. B. The psychophysics of climate. American Journal of Psychology, 1909, 20, 1-14.

Toops, H. The criterion. Educational and Psychological Measurement, 1944, 4, 271-297.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Travers, R. M. W. A critical review of the validity and rationale of the forced-choice technique. Psychological Bulletin, 1951, 48, 62-70.

Trawick, M., & Munger, A. M. Objective criteria of service station dealer success. Social science research reports, volume III. Personnel review and evaluation. New York: Standard Oil of New Jersey, 1962.

Tucker, M. F., Cline, V. B., & Schmitt, J. R. Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. Journal of Applied Psychology, 1967, 51, 131-138.

Turner, W. W. Dimensions of foreman performance: A factor analysis of criterion measures. Journal of Applied Psychology, 1960, 44, 216-223.

Uhrbrock, R. S. Standardization of 724 rating scale statements. Personnel Psychology, 1950, 3, 285-316.

Uhrbrock, R. S. 2000 scaled items. Personnel Psychology, 1961, 14, 375-420.

Van Dusen, A. C. Importance of criteria in selection and training. Educational and Psychological Measurement, 1947, 7, 498-504.

Vielhaber, D. P., & Gotthiel, E. First impressions and subsequent ratings of performance. Psychological Reports, 1965, 17, 916.

Viteles, M. S. Industrial psychology. New York: W. W. Norton, 1932.

Viteles, M. S. A dynamic criterion. Occupations, 1936, 14, 963-967.

Viteles, M. S. The aircraft pilot: 5 years of research, a summary of outcomes. Psychological Bulletin, 1945, 42, 489-526.

Viteles, M. S. Motivation and morale in industry. New York: W. W. Norton, 1953.

Volkmann, J. The method of single stimuli. American Journal of Psychology, 1932, 44, 808-809.

Vroom, V. H. Projection, negation and the self-concept. Human Relations, 1959, 12, 335-344.

Vroom, V. H. Work and motivation. New York: Wiley, 1964.

Vroom, V. H. Industrial social psychology. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (2nd ed.) Volume 5. Applied social psychology. Reading, Mass.: Addison-Wesley, 1969.

Vroom, V. H. Leadership. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.

Vroom, V. H., & Yetton, P. W. Leadership and decision-making. Pittsburgh: University of Pittsburgh Press, 1973.

Wallace, S. R. Criteria for what? American Psychologist, 1965, 20, 411-417.

Weiss, W. Effects of unbalanced response scales on judgments of social stimuli. Psychological Reports, 1963, 12, 403-414.

Weitz, J. Criteria for criteria. American Psychologist, 1961, 16, 208-213.

Weitz, J. The use of criterional measures. Psychological Reports, 1964, 14, 803-817.

Weitz, J., & Nuckols, R. C. A validation study of "How Supervise?" Journal of Applied Psychology, 1953, 37, 7-8.

Wells, F. L. A statistical study of literary merit. Archives of Psychology in New York, 1907, 1, No. 7.

Wever, E. G., & Zener, K. E. The method of absolute judgment in psychophysics. Psychological Review, 1928, 35, 466-493.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. Training interviewers to eliminate contrast effects in employment interviews. Journal of Applied Psychology, 1973, 57, 233-236.

Wexley, K. N., Yukl, G. A., Kovacks, S. Z., & Sanders, R. E. Importance of contrast effects in employment interviews. Journal of Applied Psychology, 1972, 56, 45-48.

Wherry, R. J. The control of bias in ratings: VII. A theory of rating (PRB Report No. 922). Washington, D.C.: Personnel Research Branch, Department of the Army, 1952.

Wherry, R. J. The past and future of criterion evaluation. Personnel Psychology, 1957, 10, 1-5.

Whitlock, G. H. Application of the psychophysical law to performance evaluation. Journal of Applied Psychology, 1963, 47, 15-23.

Whitlock, G. H., Clouse, R. J., & Spencer, W. F. Predicting accident proneness. Personnel Psychology, 1963, 16, 35-44.

Whyte, W. F. Money and motivation. New York: Harper, 1955.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison-Wesley, 1973.

Wiley, L. Relation of characteristics ratings to performance ratings. Journal of Industrial Psychology, 1964, 2, 7-15.

Wiley, L., & Jenkins, W. S. Selecting competent raters. Journal of Applied Psychology, 1964, 48, 215-217.

Williams, W. E., & Seiler, D. A. Relationship between measures of effort and job performance. Journal of Applied Psychology, 1973, 57, 49-54.

Willingham, W. W., & Jones, M. B. On the identification of halo through analysis of variance. Educational and Psychological Measurement, 1958, 18, 403-407.

Wood, M. T. Power relationships and group decision making in organizations. Psychological Bulletin, 1973, 79, 280-293.

Zavala, A. Development of the forced-choice rating scale technique. Psychological Bulletin, 1965, 63, 117-124.

Zedeck, S., & Baker, H. T. Nursing performance as measured by behavioral expectation scales: A multitrait-multirater analysis. Organizational Behavior and Human Performance, 1972, 7, 457-466.

Zedeck, S., Imparato, N., Krausz, M., & Oleno, T. Development of behaviorally anchored rating scales as a function of organizational level. Journal of Applied Psychology, 1974, 59, 249-252.

Zedeck, S., Jacobs, R., & Kafry, D. Behavioral expectations: Development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology, 1976, 61, 112-115.

Zedeck, S., Kafry, D., & Jacobs, R. Format and scoring variations in behavioral expectation evaluations. Organizational Behavior and Human Performance, 1976, 17, 171-184.

VITA

WILLIAM I. SAUSER, JR.

Major Field: Industrial/Organizational Psychology

Address: Department of Psychology
Auburn University
Auburn, AL 36830
Telephone: (205) 826-4412

Personal Information:

Date and place of birth: June 8, 1950
Jackson, Mississippi

Marital status: Married, no dependents

Academic History:

9/68 - 3/72. Georgia Institute of Technology. B.S. in Behavioral Management (Cum Laude). Minors: Social Science, Psychology.

3/72 - 12/74. Georgia Institute of Technology. M.S. in Psychology. Major: Industrial/Organizational Psychology. Minor: Personnel Management and Labor Relations.

1/75 - 12/78. Georgia Institute of Technology. Ph.D. in Psychology. Major: Industrial/Organizational Psychology. Minor: Personnel Management and Labor Relations.

9/75 - 3/76. Georgia State University. Advanced graduate course work in psychodiagnostics and individual appraisal.

Master's Thesis: Sex Differences in Job Satisfaction (Cited by the Georgia Tech chapter of Sigma Xi as the Outstanding Master's Thesis in Science for 1975.) (Thesis director: Dr. C. M. York)

Doctoral Dissertation: A Comparative Evaluation of the Effects of Rater Participation and Rater Training on Characteristics of Employee Performance Appraisal Ratings and Related Mediating Varibles. (Dissertation director: Dr. E. H. Loveland)

Professional Experience:

10/78 - present. Vice President, Southern Professional Services, Inc. Provide consultation to industry regarding advertising and marketing strategy; employee selection, training, and development; equipment, product, and workplace design; and other business-related applications of behavioral theory and research.

9/78 - present. Assistant Professor and Coordinator, Industrial-Organizational Psychology Program, Department of Psychology, Auburn University. Same duties as below, with additional graduate-level teaching and research supervision responsibilities.

9/77 - 9/78. Instructor and Coordinator, Industrial-Organizational Psychology Program, Department of Psychology, Auburn University. Full responsibility for teaching undergraduate courses in general, industrial, design, and experimental social psychology. Additional responsibilities include coordinating the Industrial-Organizational program, serving on the Undergraduate Affairs, Registration, and Pre-College Counseling and Graduate Admissions committees, and supervising student research.

3/77 - 9/77. Graduate Teaching Assistant, School of Psychology, Georgia Institute of Technology. Full responsibility for teaching a senior-level undergraduate course in social psychology.

9/75 - 9/77. Senior Research Assistant, Southern Regional Office, Educational Testing Service. Participated in a wide variety of activities as research assistant to the Vice President and Director. Duties included: Developing and evaluating research proposals and requests for proposals; preparing research reports and professional papers and addresses; assisting in the implementation of large-scale testing programs (including supervision of test centers) and field studies; implementing survey research projects; assisting in the editing of test manuals and technical reports; developing prototype affective measures; maintaining files and depositories of test and research materials; participating in workshops and seminars; and consulting on a wide range of educational, methodological, and organizational problems.

6/75 - 9/75. Cooperative Trainee, Department of Education and Training, Lockheed-Georgia Company. Outlined (under supervision) a comprehensive proposal to select and evaluate Manufacturing Supervisor Cooperative Trainees, including job analysis, development/identification of performance measures and appropriate selection instruments, and validation studies. Supervised and participated in a multi-method analysis of the Manufacturing Supervisor's job. Methods used included direct observation, in-depth interviewing, and application of the Position Analysis Questionnaire and the Critical Incident Technique. Also provided assistance in the development of education/training programs for skilled and professional employees.

9/74 - 6/75. Instructor (Part-time), Kennesaw Junior College. Full responsibility for teaching one course in introductory psychology per quarter.

7/73 - 9/74. Graduate Research Assistant, Health Systems Research Center, Georgia Institute of Technology. Participated in the design and implementation of a multi-disciplinary organizational and behavioral evaluation of a federally-funded incentive reimbursement project in the health care delivery field. Duties included literature review, experimental design, questionnaire construction, data analysis, and report writing. Also participated in proposal development for additional HSRC research projects. Supervisor: C. M. York

6/72 - 6/73. Research Assistant, Southern Regional Education Board. Assisted in the evaluation of the Mental Health Associate degree program. Duties consisted of data tabulation and statistical analysis. Supervisor: E. J. Baker

3/72 - 6/72. Graduate Assistant, School of Psychology, Georgia Institute of Technology. Assisted in the preparation of a content-valid comprehensive departmental examination in introductory psychology. Duties included collecting and assembling objective questions and categorizing them in terms of content and skill domains.

Publications:

Sauser, W. I., Jr., & York, C. M. Sex differences in job satisfaction: A re-examination. Personnel Psychology, 1978, 31, 537-547.

Sauser, W. I., Jr., Arauz, C. G., & Chambers, R. M. Exploring the relationship between level of office noise and salary recommendations: A preliminary research note. Journal of Management, 1978, 4, 57-63.

Anderson, S. B., & Sauser, W. I., Jr. Measurement of test anxiety. Manuscript submitted for publication, 1978.

Technical Reports:

Emerzian, A. D. J., Harrison, B. K., Hollis, M. T., Landry, R. A., Sauser, W. I., Jr., & York, C. M. Research outline for the behavioral analysis of the Group Reimbursement Incentive Project being conducted by the Birmingham Regional Hospital Council (Social Security Administration contract SSA-PMB-73-154). Atlanta: Health Systems Research Center, Georgia Institute of Technology, November 1973.

Emerzian, A. D. J., Harrison, B. K., Hollis, M. T., Landry, R. A., Sauser, W. I., Jr., & York, C. M. Evaluation of the Group Reimbursement Incentive Project being conducted by the Birmingham Regional Hospital Council: Semiannual progress report #2 (Social Security Administration contract SSA-PMB-73-154). Atlanta: Health Systems Research Center, Georgia Institute of Technology, March 1974.

Emerzian, A. D. J., Harrison, B. K., Hollis, M. T., Landry, R. A., Sauser, W. I., Jr., & York, C. M. Evaluation of the Group Reimbursement Incentive Project being conducted by the Birmingham Regional Hospital Council: A Supplement to semiannual progress report #2--A revised research design (Social Security Administration contract SSA-PMB-73-154). Atlanta: Health Systems Research Center, Georgia Institute of Technology, July 1974.

Emerzian, A. D. J., Hollis, M. T., & Sauser, W. I., Jr. Housekeeping employee performance appraisal system: A rationale and procedure. Atlanta: Health Systems Research Center, Georgia Institute of Technology, July 1974.

Sauser, W. I., Jr., & York, C. M. Employee opinion research: Department of Revenue, State of Georgia. Atlanta: School of Psychology, Georgia Institute of Technology, September 1974.

Emerzian, A. D. J., Harrison, B. K., Hollis, M. T., Landry, R. A., & Sauser, W. I., Jr. Evaluation of the Group Reimbursement Incentive Project being conducted by the Birmingham Regional Hospital Council: Semiannual progress report #3 (Social Security Administration contract SSA-PMB-73-154). Atlanta: Health Systems Research Center, Georgia Institute of Technology, September 1974.

Presentations to Professional Societies:

Sauser, W. I., Jr. Dimensions of teacher behavior. Paper presented at the meeting of the Georgia Psychological Association, Savannah, GA, May 1974.

Sauser, W. I., Jr. Sex differences in job satisfaction. Paper presented at the meeting of the Southeastern Psychological Association, Atlanta, March 1975. (A revision of this paper is published in Personnel Psychology, 1978, 31, 537-547.)

Sauser, W. I., Jr., & York, C. M. Job satisfaction of Georgia state government employees. Paper presented at the meeting of the Georgia Psychological Association, Atlanta, May 1975.

Arauz, C. G., & Sauser, W. I., Jr. The effects of office noise upon interpersonal judgments made in a personnel manager simulation. Paper presented at the meeting of the Georgia Psychological Association, Atlanta, May 1975. (A revision of this paper is published in Journal of Management, 1978, 4, 57-63.)

Anderson, S. B., & Sauser, W. I., Jr. Are psychologists qualified to evaluate education/training programs? In S. B. Anderson (Chair), The psychologist as program evaluator. Symposium presented at the meeting of the Southeastern Psychological Association, New Orleans, March 1976. (A revision of this paper is published as Chapter Nine: "Training Evaluators and Evaluating Their Competencies," in Anderson, S. B., & Ball, S., The profession and practice of program evaluation. San Francisco: Jossey-Bass, 1978.)

Sauser, W. I., Jr. Planning for job analysis. In W. I. Sauser, Jr. & C. G. Arauz (Chair), Job analysis in personnel decision-making. Symposium presented at the meeting of the Georgia Psychological Association, Columbus, GA, May 1976.

Sauser, W. I., Jr., & Sauser, L. D. Sex discrimination in terms of pay in a department of state government. Paper presented at the meeting of the Georgia Psychological Association, Columbus, GA, May 1976.

Sauser, W. I., Jr. Observation and in-depth interviewing. In W. I. Sauser, Jr., & C. G. Arauz (Chair), Job analysis methods in application. Symposium presented at the meeting of the Southeastern Psychological Association, Hollywood, FL, May 1977.

Sauser, W. I., Jr. The Position Analysis Questionnaire. In W. I. Sauser, Jr., & C. G. Arauz (Chair), Job analysis methods in application. Symposium presented at the meeting of the Southeastern Psychological Association, Hollywood, FL, May 1977.

Anderson, S. B., & Sauser, W. I., Jr. Measurement of test anxiety and worry. In S. B. Anderson (Chair), Test anxiety revisited. Symposium presented at the meeting of the Southeastern Psychological Association, Atlanta, May 1977. (A revision of this paper was submitted for publication, 1978.)

Sauser, W. I., Jr. Behaviorally anchored rating scales. In W. I. Sauser, Jr. (Chair), Methods for employee performance appraisal. Symposium presented at the meeting of the Georgia Psychological Association, Atlanta, May 1977.

Sauser, W. I., Jr. A theoretical model of academic dishonesty. In W. I. Sauser, Jr., & B. G. Witmer (Chair), Cheating and lying: Determinants of dishonesty. Symposium presented at the meeting of the Georgia Psychological Association, Atlanta, May 1977.

Consulting Experience:

Department of Revenue, State Government of Georgia (1974; Dr. C. M. York, Project Director). Project consisted of assessing employee morale and attitudes toward specific job aspects. Methods for improving morale were suggested based on survey results.

Lockheed-Georgia Company (1975; Dr. E. H. Loveland, Project Director). Project involved developing selection, training, and performance evaluation methods for first-line aircraft manufacturing supervisors. Continued informal involvement through 1977.

Georgia Power Company (1976; Dr. C. M. York, Project Director). Project involved assessing consumer attitudes toward the implementation of metering systems designed to reinforce the conservation of electricity through non-peak time usage.

General Assembly, Presbyterian Church in the United States (1977; Dr. C. M. York, Project Director). Project consisted of evaluating numerous General Assembly programs through the use of questionnaires administered to representative samples of several of the Church's major constituencies.

Alabama State Nurses' Association (1978; with Drs. A. A. Armenakis and S. B. Green). Projects involve developing scales to evaluate the performance of the expanded-role nurse; measuring communication among and role definition of hierarchical levels of nursing in hospitals.

RLC Electronics, Inc. (1978; with Dr. C. W. Jenkins). Projects involve preparing and implementing market research surveys for a variety of commercial products.

Continuing Education Workshops Conducted:

Personnel Relations: Performance Appraisal. Searcy Hospital, Mt. Vernon, Alabama, September 14-15, 1978. (Sponsored by the Office of Public Service and Research, School of Arts and Sciences, Auburn University.)

Personnel Relations: Performance Appraisal. Lurleen B. Wallace Developmental Center, Decatur, Alabama, September 29, 1978. (Sponsored by the Office of Public Service and Research, School of Arts and Sciences, Auburn University.)

Memberships in Professional Organizations:

Alabama Psychological Association
Alpha Kappa Psi (professional business fraternity)
American Psychological Association (student affiliate)
Beta Gamma Sigma (business honor society)
Sigma Xi (associate member)
Southeastern Industrial/Organizational Psychologists Association
Southeastern Psychological Association
Southern Society for Philosophy and Psychology (associate member)

Other Professional Activities:

Occasional reviewer, Journal of Management.